Real-Time Elbow Fracture Detection on Mobile Devices: Performance and Limitations

Filip Sosnowski * and Bujar Raufi *

Posted Date: 26 February 2026

doi: 10.20944/preprints202511.0873.v2

Keywords: YOLOv11; mobile artificial intelligence; fracture detection; domain adaptation; edge computing; computer vision

 doi:10.20944/preprints202511.0873.v2

Article

Real-Time Elbow Fracture Detection on Mobile Devices: Performance and Limitations

Filip Sosnowski and Bujar Raufi *

Technological University Dublin (TU Dublin), Ireland
* Correspondence: bujar.raufi@tudublin.ie

Abstract

This study investigates the feasibility and performance of deploying YOLOv11-based elbow fracture detection models on mobile devices for real-time clinical diagnosis. Motivated by clinicians' poor diagnostic accuracy (54.4%) in interpreting elbow radiographs, we developed and evaluated a smartphone application capable of performing inference via three pathways: photo library analysis, direct camera capture, and live video streaming. Two YOLOv11 models were trained on approximately 1,100 elbow X-ray images using systematic hyperparameter optimisation and deployed in both $FP16$ and $FP32$ quantisation formats. On digital radiographs, Model 1 achieved strong performance with 93.4% accuracy, 92.7% F1-score, and 69.3% mAP@50, demonstrating computational efficiency with RAM usage below 0.29GB and CPU consumption under 25%. However, substantial performance degradation occurred during real-world camera-based testing, with F1-scores declining to $31 - 60.3\%$ for static photographs and $28.8 - 43.1\%$ for live detection. Independent validation on 214 external images yielded moderate classification accuracy ($64 - 66\%$), highlighting generalisation challenges that may stem from limited dataset diversity. The study demonstrates that while YOLOv11 achieves clinically relevant accuracy on curated digital radiographs with mobile-friendly computational requirements, current training paradigms prove insufficient for robust camera-based deployment. Severe domain shift effects, environmental sensitivity, and class imbalance issues (fracture detection F1: $69.2 - 81.6\%$ versus non-fracture: $93.7 - 94\%$) represent critical barriers to clinical implementation. These findings emphasise that exciting laboratory metrics do not guarantee real-world utility and establish that safe clinical deployment requires diverse datasets incorporating varied acquisition conditions, environmental compensation strategies, and enhanced fracture detection sensitivity to meet patient safety standards.

Keywords: YOLOv11; mobile artificial intelligence; fracture detection; domain adaptation; edge computing; computer vision

1. Introduction

Recent developments in artificial intelligence (AI), particularly deep learning, have shown significant potential in assisting clinicians with bone fracture detection and diagnosis. The research indicates that AI models, especially those using convolutional neural networks (CNNs), achieve diagnostic accuracy, sensitivity, and specificity comparable to or sometimes exceeding that of clinicians, suggesting AI can serve as a valuable adjunct in clinical practice, which underscores their potential impact [1], [2], [3]. Studies demonstrate that AI assistance improves clinicians' sensitivity in detecting fractures, reduces diagnostic errors, and shortens reading times, thereby enhancing both the speed and accuracy of diagnosis in emergency and routine settings [4],[5],[3].

AI systems have proven effective across various imaging modalities (X-ray, CT, MRI) and body regions. Commercial AI algorithms have performed well in real-world emergency department scenarios [6]. However, challenges such as clinical translation, regulatory approval, and ensuring model robustness and safety remain significant barriers to widespread adoption [4], [2]. AI holds great promise for

 doi:10.20944/preprints202511.0873.v2

improving fracture detection workflows and patient outcomes, but further work is needed to address implementation barriers and optimise integration into clinical practice [1],[2],[7]. Furthermore, the ability of clinicians to interpret X-ray images of elbow injuries remains inadequate, even in developed countries, and empirical data indicate an accuracy rate of only 54.4% for the correct diagnosis of children's elbow injuries when clinicians are presented with a set of ten such images [8]. In contrast, models such as DenseNet-201 have achieved an accuracy of 94.1% in performing the same task [9].

Given that in many low-income nations access to immediate medical attention is limited, communicating with a clinician via a smartphone has been shown to be effective for obtaining a real-time diagnosis in emergency cases [10]. Research on integrating bone fracture detection models into smartphones and edge devices is currently scarce, yet it demonstrates the potential to improve healthcare accessibility [11]. Despite advancements in AI for medical imaging, generalisation remains an issue many models face. Particularly when performance declines when analysing data from patients who differ from those used to train the model. For example, an AI model for detecting cervical spine fractures demonstrated poor real-world performance despite promising results during testing [12]. Other hurdles in this area include model explainability, which is necessary for clinicians to understand the logic and reasoning underlying AI predictions in complex scenarios [13].

Rayan et al. address a critical issue in the preliminary assessment of radiographs, which is typically conducted by clinicians without radiology specialisation. These practitioners may lack the expertise to discern complex fracture patterns specific to certain demographic groups, such as children or senior citizens, especially in high-pressure settings like emergency departments and urgent care facilities, where precise, rapid patient triage is crucial. In cases where immediate consultation with a radiologist is unfeasible, implementing an AI-based fracture detection system could substantially enhance the efficacy and accuracy of patient triage [14]. Delays or the absence of treatment for elbow fractures can result in long-term issues, including immediate complications such as nerve palsy, joint immobility, and cubitus varus, a condition marked by improper alignment of the elbow joint [15]. The gravity of these conditions underscores the importance of having tools that can increase the speed and accuracy of diagnosing elbow fractures.

This article aims to identify and evaluate the feasibility, in terms of performance, accuracy, and usability, of AI models deployed on mobile devices and their application for real-time interpretation of elbow X-ray images in a clinical setting, using a Convolutional Neural Network (CNN) model. Specifically, this system targets triage support in emergency departments lacking 24/7 radiology coverage and telemedicine consultation in resource-limited settings, where high sensitivity for fracture detection is prioritized to minimize missed diagnoses requiring urgent intervention.

This article is structured as follows: In Section 2, we review prior work focused on defining and utilising fracture detection methods for mobile devices. Section 3 details the experimental setup for real-time detection of fractures. Section 4 presents the results, including a critical discussion. Finally, Section 6 concludes the article, offering suggestions for future research.

2. Related Work

Medical imaging technologies, including MRI, X-ray, and PET scans, are essential to modern diagnostic healthcare. Although critical, clinician interpretations can be affected by fatigue, cognitive biases, and proficiency variability, leading to inconsistent diagnoses [16]. Artificial intelligence addresses these challenges by enhancing diagnostic precision through its pattern recognition abilities, detecting minute details beyond human capability, speeding up data analysis, and reducing healthcare costs by optimising resources [17],[16]. Recent AI applications in medical imaging have highlighted significant progress, particularly in cancer detection. CNNs have demonstrated efficacy in detecting brain tumours from MRI and PET scans, improving liver and pancreas imaging through accelerated segmentation and scanning, and enhancing breast cancer detection accuracy with limited training data [13].

 doi:10.20944/preprints202511.0873.v2

2.1. Bone Fracture Detection

Artificial intelligence has shown remarkable proficiency in identifying bone fractures, with CNN-based models regularly outperforming human radiologists in terms of accuracy, sensitivity, and specificity [12]. The relevance of this technology further emphasises the fact that misdiagnosed bone fractures account for roughly 80% of medical errors in emergency departments, a challenge intensified by the increasing scarcity of radiologists due to hiring delays and workforce attrition [18]. Investigations over the past twenty years indicate that AI models offer high classification accuracy, featuring an average sensitivity rate of 92%, while radiograph-based detection achieves a pooled sensitivity of 94% [2]. Several strategies have recently been pursued to enhance fracture detection models. The incorporation of attention mechanisms, which enable models to focus on crucial areas of input data, has improved detection efficiency [12]. Transfer learning, involving the refinement of models pre-trained on extensive datasets for fracture detection, has shown noteworthy accuracy with minimal need for specialised data [19]. Recent research applying YOLOv8 for the quality control of elbow radiographs has demonstrated the methodology's high performance in terms of evaluation time, precision, and recall, as well as its capability to identify distinct elbow joint elements [9]. An analysis of detection frameworks highlights significant trade-offs. One-stage frameworks such as YOLO, which operate as Single Shot Detector (SSD), require considerably fewer computational resources and offer greater speed than two-stage frameworks, albeit at the cost of reduced accuracy. Although specific studies claim that YOLO is unmatched in speed, comprehensive reviews present a nuanced view, suggesting that YOLO is ideal for speed-focused applications, while recommending models such as Faster R-CNN or RetinaNet for high accuracy and effective detection of small objects. CT scans are shown to be more accurate than X-ray images, achieving up to 100% sensitivity in specific scenarios, compared to the 75.2% observed with AI-aided X-ray fracture detection. Additionally, an evaluation of BoneView, a commercially available AI diagnostic tool, indicated that it enhanced radiologists' sensitivity and negative predictive value for wrist and hand fracture detection by an average of 5.3% at the patient level, enabling junior radiologists to match the diagnosis accuracy of their senior peers when utilising the tool [18].

2.2. Elbow Fracture Challenges

Bone complexity and specificity adapted to a patient contribute to up to 11% of acute fractures being missed by physicians in emergency settings, compared with those missed by trained radiologists. The situation is particularly problematic in high-volume emergency departments and urgent care centres where radiologists may not be readily available, making accurate and efficient patient triage paramount [20]. Concealed fractures pose a significant diagnostic challenge, as they may only be visible from specific angles or X-ray views. The relatively smaller size of bones in children and smaller individuals reduces image clarity. Models using DenseNet-201 could detect such fractures with 94.1% accuracy and 98.7% AUC during training, achieving 90.5% and 89.3% accuracy in external validation—superior to other compared models [9]. The clinical importance of accurate diagnosis is emphasised by potential long-term consequences of missed or delayed treatment, including nerve palsy, joint stiffness, and cubitus varus (elbow joint misalignment) [18]. Meta-analysis of six studies examining deep learning models for elbow injury detection revealed high performance metrics with pooled sensitivity of 0.93, specificity of 0.89, and AUC of 0.95, suggesting these models could reliably screen patients in busy emergency settings [15].

The importance of models capable of generalising across different images and real clinical settings is paramount. Failure patterns in elbow fracture detection can lead to delayed or unnecessary treatment, but most errors do not result in significant long-term harm if promptly identified and managed. For example, missed elbow fractures (false negatives) can result in delayed immobilisation or surgery, potentially increasing the risk of complications such as joint instability, malunion, or chronic pain. However, studies show that when misdiagnoses are identified and corrected within a day, no additional morbidity or need for surgical intervention occurs, especially with routine radiologist review. Most

 doi:10.20944/preprints202511.0873.v2

false negatives are minor or occult fractures that often do not require a change in treatment. Nonetheless, in rare cases (e.g., olecranon or supracondylar fractures), closer follow-up may be needed to avoid adverse outcomes.[21,22] False positive, on the other hand, can lead to overdiagnosis, unnecessary immobilisation, follow-up imaging, or even surgical interventions, causing patient anxiety, increased healthcare costs, and potential iatrogenic harm[22–24]. This may burden emergency departments and radiology services, especially in cases where interpretation is more challenging [23,24].

2.3. Current Approaches

In recent years, various artificial intelligence (AI) methods and strategies have been explored to enhance the identification of elbow fractures. One of these approaches leverages two distinct deep convolutional neural network (CNN) models that generate heatmaps to assist clinicians, utilising an extensive dataset comprising 1,956 elbow x-rays, assessed by a panel of eight radiologists. The study observed that the efficacy of AI models in enhancing or impairing diagnostic performance is contingent upon the specific model employed [25]. DenseNet-201, for instance, was implemented for diagnosing elbow fractures, achieving an accuracy of 94.1% [9]. Nevertheless, this model demonstrated lower precision and recall relative to others, with VGG16 emerging as the superior performer among the examined deep convolutional neural networks (DCNNs) [26]. Recent advancements have extensively investigated YOLO variants for fracture detection. For instance, the application of YOLOv8 to wrist injuries attained a precision of 77.80% [27]. In other studies, the introduction of ghost convolutions to YOLOv11 facilitated an inference time of 2.4 ms, achieving a mean Average Precision (mAP) of 53.5% at an Intersection over Union (IoU) threshold of 0.5 on the GRAZPEDWRI-DX dataset [28]. However, evaluations of YOLOv9, replicated in [9], reported mAP scores of 65.46% at IoU=50 and 43.73% at IoU=50-95, challenging previous benchmark assertions [28]. These investigations highlight the profound impact of dataset quality and diversity on detection models, with significant implications arising from the limited representation of bone anomalies and soft tissue data in the GRAZPEDWRI-DX dataset. Altmann-Schneider et al.'s evaluation of the BoneView algorithm for the most prevalent fractures (forearm, elbow, lower leg) included 1,000 radiographs per body part, for elbow radiographs, where the average patient age was 7.7±3.7 years with a gender ratio of 55% male to 45% female, classifying dubious cases (50-90% confidence) as fractures achieved a sensitivity of 91.5% but specificity of 63.7%. When these were classified as non-fractures, sensitivity decreased to 80.5%, while specificity increased to 94.9%. This classification inconsistency underlined the need for further refinement for dependable clinical implementation [29]. Dupuis et al. validated Milvue's AI algorithm, observing a negative predictive value of 92%. However, they identified methodological shortcomings, as the assessment focused on agreement with radiologists rather than genuine diagnostic performance [20]. Multiple studies emphasise the pivotal role of high-quality, diverse datasets, asserting their necessity for optimising model performance. Kutbi illustrated that high-quality annotated training sets—although currently rare and costly to develop are essential for producing robust models across diverse populations and imaging modalities [2]. A study by Zech et al. further corroborates this, showing that an open-source fracture detection algorithm in children (childfx.com) can substantially improve diagnostic accuracy for both physicians and radiologists, increasing physician sensitivity from 0.842 to 0.858 and radiologist sensitivity from 0.781 to 0.883. The findings suggest that AI integration predominantly bolsters the performance of less specialised and experienced interpreters, bringing their diagnostic capabilities closer to those of subspecialists [30].

2.4. Application in Mobile and Edge Environments

Despite extensive research on fracture detection, few studies address integrating smartphones and edge devices [31]. MobileNet appears to be the most researched in this domain, with a CNN architecture designed for mobile devices that minimises parameters to accelerate computation at the cost of reduced efficiency compared to other deep CNNs [32]. Studies involved an analysis of MobileNetV2 for bone fracture detection using MobileNetV2 tested alongside other CNN architectures such as VGG16, ResNeXt, AlexNet and SFNet. The results are inconclusive, showing conflicting

 doi:10.20944/preprints202511.0873.v2

results for MobileNetV2, with a range of 59% F1 score to 48% accuracy, compared to models that achieve 93.5% accuracy as reported in [32] and [33]. Furthermore, a comparison of MobileNetV3 with ResNet50 using the MURA dataset (20,335 images across elbows, hands, and shoulders) found that MobileNetV3 achieved 78.37% accuracy for elbow fractures, versus ResNet50's 76.83%, indicating that MobileNetV3's fast inference times make it viable for real-time clinical applications on smartphones and tablets [31].

Although the research using non-MobileNet object detection models for bone fracture detection from radiographs on edge devices is generally lacking, a comparison of MobileNetV2 against YOLO-X for obscured facial recognition found that YOLO-X outperformed across accuracy, precision, recall, and F1 score, suggesting YOLO could be a suitable candidate for mobile fracture detection deployment [34].

Despite AI's demonstrated capabilities, high-quality, diverse datasets accounting for varied populations, rare fracture types, different viewing angles, lighting conditions, and camera distances remain scarce. The quality of the dataset fundamentally determines model performance and its real-world applicability. While CNNs show promise for clinical elbow X-ray diagnosis, research lacks evaluation of system requirements, resource usage, and actual smartphone and edge efficiency for these models. The feasibility of smartphone and edge applications for real-time clinical assistance remains underexplored, representing a significant opportunity to improve healthcare accessibility in resource-limited settings and emergency departments.

The novelty introduced in this paper is that it outlines a first systematic end-to-end study that trains, optimises, quantises, and deploys a modern YOLOv11 object-detection model directly onto a consumer smartphone. It also evaluates not only diagnostic accuracy on digital elbow radiographs but also real-world mobile use cases, including photo capture of printed radiographs and live camera streaming, along with detailed CPU/RAM profiling on low and high-end devices. Most prior fracture-detection studies, which focus on offline classification accuracy using heavy CNNs (e.g., DenseNet, ResNet) on curated PACS images, this work explicitly addresses mobile feasibility, latency, quantisation effects (FP16 vs. FP32), and domain shift introduced by camera acquisition. In contrast to earlier mobile AI efforts that rely primarily on lightweight classifiers such as MobileNet and report mixed diagnostic performance, this study demonstrates that a single-stage detector (YOLOv11) can achieve competitive accuracy on digital radiographs while remaining computationally efficient enough for real-time edge deployment, thereby unifying localisation, classification, and inference within a single mobile pipeline.

3. Materials and Methods

This section elaborates on and analyses the design decisions made in the experimental setup. The experiment is structured into four phases, with each focusing on distinct elements of the design. The initial phase encompasses data preparation and model training methodologies, while the subsequent phase involves exporting the Yolo model and developing the application. The third phase is dedicated to implementing particular app features, as well as the associated evaluation and testing. The overall pipeline is outlined in Figure 1. This systematic workflow ensures robust model development, efficient deployment, and thorough validation of the fracture detection system.

 doi:10.20944/preprints202511.0873.v2

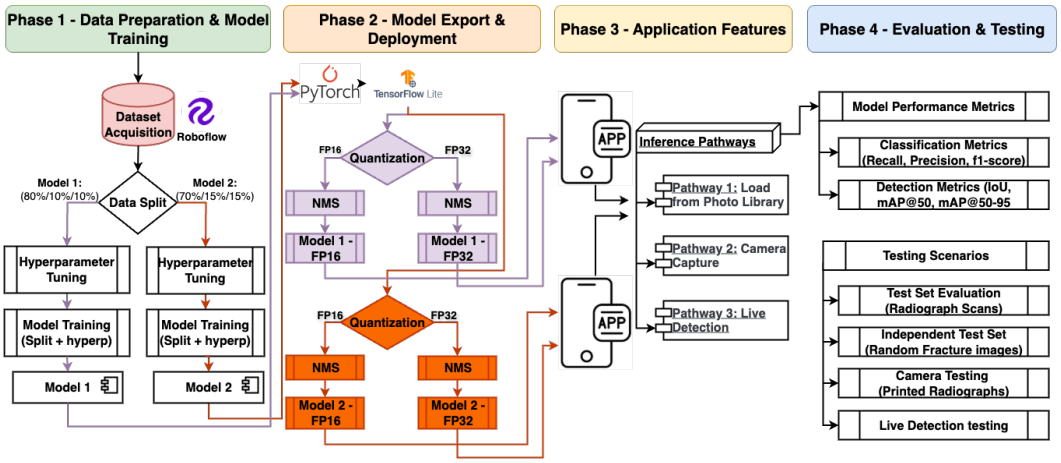


Figure 1. End-to-end pipeline for elbow fracture detection in radiographic images. The pipeline comprises four distinct phases: Phase 1 (Data Preparation & Model Training) involves dataset acquisition via Roboflow, 60/20/20 train/validation/test splitting, hyperparameter tuning, and parallel training of two model variants. Phase 2 (Model Export & Deployment) includes model quantization using PyTorch and TensorFlow Lite frameworks, with both $FP16$ and $FP32$ precision variants, followed by neural network model size (NMS) optimization, producing four deployable models (Model 1-FP16, Model 1-FP32, Model 2-FP16, Model 2-FP32). Phase 3 (Application Features) demonstrates three inference pathways through a mobile application interface: loading images from the photo library, real-time camera capture, and live fracture detection. Phase 4 (Evaluation & Testing) encompasses comprehensive model assessment including classification metrics ($recall$, $precision$ and $f1-score$), detection performance metrics (IoU, $mAP@50$, $mAP@50-95$), multiple testing scenarios with radiograph scans, independent validation using random fracture images, camera-based testing with printed radiographs, and live detection capability verification.

3.1. Data Preprocessing and Augmentation

The foundation of the entire fracture detection system is built on careful dataset preparation and systematic model training. The process begins with acquiring a labelled dataset from Roboflow containing approximately 1,100 adult elbow X-ray images. This dataset includes YOLO-formatted annotations with bounding box coordinates and binary classifications ($fractured$ vs. $non-fractured$). The dataset comprised 367 fractured and 733 non-fractured elbow images, yielding a 2:1 class imbalance ratio favouring non-fractures. For Model 1 (80/10/10 split): training contained 293 fractured / 586 non-fractured images; for Model 2 (70/15/15 split): 256 fractured / 513 non-fractured. A significant limitation of the Roboflow dataset is the absence of comprehensive clinical metadata, which prevented verification of key dataset characteristics. Specifically, we could not ascertain the fracture subtype distribution (e.g., supracondylar, lateral condyle, radial head), the radiographic view composition (anteroposterior, lateral, oblique), the imaging acquisition protocols (equipment manufacturers, exposure parameters), or the annotation reliability metrics (inter-rater agreement for bounding box coordinates). Reported metrics represent aggregate performance across heterogeneous fracture presentations; clinical effectiveness may vary substantially by fracture complexity, with occult fractures potentially underdetected relative to displaced fractures. Future work should stratify performance by validated fracture classification systems (AO/OTA). This metadata gap represents a substantial threat to external validity, as performance estimates may not generalize across diverse fracture presentations, pediatric versus adult populations, or varied institutional imaging standards. The supplementary materials regarding the dataset are provided to facilitate reproducibility of the study, since the dataset lacks the necessary metadata to ensure its provenance. All radiograph images in the dataset underwent standardised preprocessing before model training. Each image is subjected to automatic contrast enhancement through contrast stretching to normalise intensity distributions across the dataset. To augment training data and improve model generalisation, each source image was duplicated with a 50% probability of horizontal flip, effectively doubling the dataset size while introducing spatial invariance. Images are resized to YOLOv11's default input resolution of 640×640 pixels with letterbox padding to maintain

 doi:10.20944/preprints202511.0873.v2

aspect ratios. Grayscale X-ray images were converted to 3-channel format through channel replication to match the model's RGB input requirements. No additional preprocessing was applied to images during inference, whether from digital radiographs or camera-captured photographs. Notably, the absence of domain-specific preprocessing for camera-acquired images (such as perspective correction, glare removal, or background segmentation) represents a deliberate design choice to evaluate the model's raw generalisation capability. However, this likely contributed to the observed performance degradation in real-world camera testing scenarios documented in Sections 4.2 and 4.3.

3.2. Model Training

Experiments were conducted on GPU P1000 model with 16GB of VRAM on a single CPU with 30GB of system RAM. Each model required approximately 8-10 hours for complete Optuna optimization (60 trials × 1,000 epochs). Convergence occurred by epoch 50-70] for Model 1 and Model 2 as evidenced by plateau in validation mAP@50 (Appendix Figures A1-A2. No explicit early stopping was applied; all trials completed 1,000 epochs to ensure thorough hyperparameter exploration. A critical step involves reorganising the dataset into two split configurations: an 80/10/10% and a 70/15/15% distribution across training, validation, and testing subsets, respectively. The dual-split strategy enables a comparative analysis of how different training/validation/testing ratios affect model performance, which is particularly important given the relatively small dataset. The decision to use fixed splits rather than k-fold CV was primarily driven by computational constraints (each model required 60 Optuna trials over 1,000 epochs, totalling approximately 120,000 training iterations per model), and the exploratory nature of this feasibility study focused on establishing baseline performance and identifying deployment challenges. However, we recognise that k-fold CV would provide more robust performance estimates with confidence intervals, reduce variance from random splits, and better utilise limited data and methodological improvements, which are explicitly recommended for future work. The training methodology emphasises rigorous hyperparameter optimisation using Optuna, an open-source framework that automates the search for optimal model configurations. Over 60 trials, the system explores combinations of five critical hyperparameters outlined in Table 1.

Table 1. Search parameters for hyperparameter tuning of CNN models

Parameters	Values
Initial Learning Rate (lr0)	[0.00001 - 0.1]
Final Learning Rate (lrf)	[0.01 - 1.0]
Batch Size	16, 32 and 64
Box Loss Weight	[0.02 - 0.2]
Classification loss weight	[0.2 - 4.0]

The hyperparameter search space design on Table 2 reflects standard practices for YOLO architectures and object detection tasks, with ranges chosen to balance computational tractability with exploration of parameter regions known to impact detection performance in medical imaging contexts. The learning rate ranges ($lr0 : 0.00001 - 0.1, lrf : 0.01 - 1.0$) span several orders of magnitude to accommodate both aggressive and conservative optimisation strategies, which is particularly important given the relatively small dataset ($1,100$ images) where optimal convergence rates are unknown a priori and too high a learning rate risks overshooting minima in sparse data scenarios. In contrast, excessively low rates may trap optimisation in local minima or require prohibitive training time within the 1,000-epoch budget. The batch sizes (16, 32, and 64) represent discrete, hardware-constrained choices that directly affect gradient estimation quality and memory usage. These specific values align with typical GPU memory limitations and the dataset size, ensuring sufficient batches per epoch ($1,100 images/64 = 17$ batches minimum). The loss weight ranges (box loss: $0.02 - 0.2$, classification loss: $0.2 - 4.0$) are asymmetric by design, reflecting YOLOv11's architectural emphasis where classification confidence typically requires stronger penalisation than bounding box coordinate regression to prevent the classifier from making overconfident predictions on ambiguous medical images. The 20-fold range for classification

 doi:10.20944/preprints202511.0873.v2

loss ($0.2 - 4.0$) versus 10-fold for box loss ($0.02 - 0.2$) acknowledges that fracture presence/absence determination is the clinically critical decision requiring robust optimisation.

YOLOv11 was selected over alternative mobile-friendly architectures based on recent evidence demonstrating superior performance for medical fracture detection. YOLOv11l achieves 96.5% precision, 95.7% recall, and 96.1% F1-score for distal radius fracture detection, outperforming both YOLOv8 variants and Faster R-CNN, while lightweight G-YOLOv11 demonstrates inference times of 2.4 ms with mAP@0.5 of 0.535 on wrist fractures[35]. Compared to mobile alternatives, YOLOv11m achieves 22% fewer parameters than YOLOv8m while maintaining higher mAP scores, and YOLO generally outperforms MobileNet SSD in accuracy, especially for complex scenes or smaller objects. YOLOv11's architectural innovations, including C3k2 blocks for improved multi-scale feature extraction and C2PSA attention modules, enhance the detection of small, occluded, or irregular objects, which is critical for identifying subtle elbow fractures.

We employ object detection (YOLOv11) rather than image-level classification for three reasons. Firstly, spatial localisation via bounding boxes enables clinicians to verify the anatomical fracture location, providing explainability that is absent in pure classification models. Secondly, the clinical workflow integration, which highlights the region-of-interest, assists non-specialist clinicians in focused examination. Finally, the object detection enables future extensibility to multi-fracture detection and fracture subtype classification. While binary classification might achieve comparable accuracy at lower computational cost, spatial information is clinically valuable for decision-support applications.

In medical imaging applications, YOLOv11 achieved state-of-the-art results in bone tumour detection at approximately 81 FPS, demonstrating that it strikes an optimal balance between clinical-grade accuracy and computational efficiency suitable for smartphone deployment.[36] This phase ultimately produces two distinct models (Model 1 and Model 2) trained on different data splits, enabling empirical comparison of how dataset allocation affects real-world performance in subsequent testing phases.

3.3. Model Export and Deployment

Phase 2 represents a critical transition from research-grade model training to practical mobile deployment, during which the trained YOLOv11 models is converted to a format optimised for resource-constrained smartphone environments. TensorFlow Lite (TFLite) is chosen as the deployment format after evaluating three potential options: PyTorch (with ExecuTorch), ONNX Runtime, and TFLite with LiteRT. This decision was informed by TFLite's superior performance, its smaller size, faster execution, and better hardware acceleration compared to alternatives. The export process, executed via Ultralytics, transformed PyTorch-trained models into mobile-optimised TFLite versions. Critically, the FP16 quantisation was enabled during export, which compresses model weights from 32-bit to 16-bit floating points, reducing model size by approximately 50% while maintaining acceptable accuracy, a crucial optimisation for mobile devices with limited storage and memory.

The crucial step in this phase is to properly understand and map the output from a trained YOLOv11 model, which utilises Non-Maximum Suppression (NMS) for object detection during model training. Since detection models often output a large number of possible detections, this results in many overlapping bounding boxes. Non-Maximum Suppression (NMS) helps to interpret these outputs by combining object detection bounding boxes that appear to belong to the same object. Subsequently, it helps remove redundant object-detection information. The NMS implemented in YOLO uses a recursive approach by doing score thresholding, i.e. removing bounding boxes with a confidence score below a specified threshold; sorting the remaining bounding boxes in descending order of confidence (sorting by confidence) and iterative selection and suppression by finding the intersection over Union (IoU) where the overlap of two bounding boxes occurs. IoU is calculated as:

$$IoU = \frac{\Xi|A \cap B|}{\Xi|A \cup B|} \tag{1}$$

 doi:10.20944/preprints202511.0873.v2

where the $\Xi|A \cap B|$ describes the area where two boxes intersect, while the $\Xi|A \cup B|$ describes the total area the two boxes occupy. The IoU of the bounding box with the highest confidence score is checked against the remaining boxes, which are then removed if they fall above a specified threshold. If other bounding boxes overlap sufficiently with the chosen bounding box, they are also removed. YOLO's default confidence threshold of 0.25 was used during evaluation, meaning detections below 25% confidence were suppressed. For clinical deployment, this threshold should be optimized via ROC analysis to achieve target sensitivity $\geq$ 95%for fracture detection, even if increasing false positive rates. Current evaluation uses detection-focused thresholds rather than safety-optimized operating points.

3.4. Application Features Implementation

The application's features are designed with clinical practicality and user accessibility in mind. The following three distinct inference pathways are implemented to accommodate different real-world usage scenarios:

1. Loading images from a photo library to enable the classification of pre-existing saved radiographs.
2. Direct camera capturing to allow for immediate fracture classification.
3. Live fracture detection that provides real-time feedback during positioning and imaging.

This multi-modal approach ensures flexibility in clinical workflows, whether reviewing archived images or conducting point-of-care assessments.

The single-page UI/UX design with colour-coded results (blue for non-fracture, red for fracture) is deliberately chosen to minimise cognitive load and enable rapid visual interpretation by healthcare providers. Large buttons and clear text enhance accessibility across diverse user groups and lighting conditions, while guideline-based responsive constraints ensure consistent functionality across various Android devices with varying screen sizes.

3.5. Evaluation and Testing

The evaluation strategy is structured to assess both technical performance and real-world applicability through progressively challenging testing scenarios. Model output processing prioritised the highest confidence prediction after Non-Maximum Suppression filtering to reduce false positives and provide clinicians with the most reliable assessment. Dual classification metrics (*precision*, *recall* and $f1 - score$) and detection metrics (IoU and mAP) are used to comprehensively evaluate the system's diagnostic accuracy and spatial localisation capability.

Precision measures the proportion of correctly predicted positive results out of all positive predictions made, and it is calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

where TP is the true positive rate and FP is the false positive rate during model prediction.

Recall is the ratio of true positive predictions to the total number of actual positive cases and is given as:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where FN is the false negative rate during model prediction.

The f1-score is the harmonic mean of precision and recall, providing a single score that balances recall and precision. The f1-score is given as:

$$f1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

The detection metric utilised is IoU (earlier explained in 3.3), along with mean average precision (mAP). The mean Average Precision serves as a common standard for assessing the efficacy of

 doi:10.20944/preprints202511.0873.v2

object detection models. The methodology for mAP calculation involves filtering predictions with 0% confidence, ranking them by confidence using NMS, and assessing each prediction's IoU against a threshold to categorise as true or false positives/negatives. Sorted predictions were used to plot the PR curve, calculating precision and recall iteratively to evaluate the object detection model's performance.

This methodology yielded a series of precision-recall pairs, facilitating the plotting of the PR curve and enabling the computation of mAP through the integration of the area beneath the curve. This was achieved using the Riemann sum:

$$mAP = \sum_{i=1}^{n} f(x_i^*)\Delta_x \tag{5}$$

where $f(x_i^*)$ is the precision value at index i, $f(x_i^*) = precision[i]$, and Δ_x denotes the recall at position $\Delta_x = recall[i] - recall[i-1]$:

Finally, the testing scenarios progressed from controlled conditions using the standard test set to an independent Roboflow[1] dataset that simulates varied image quality to physically printed radiographs under different lighting and angular orientations. The dataset is adequate, as it includes 214 additional test samples. The independent Roboflow dataset was selected for three reasons: publicly accessible provenance enabling reproducibility, sufficient sample size ($n = 214$) for meaningful statistical inference, and different institutional source than training data, providing genuine external validation. This final testing phase with printed X-rays was particularly critical, as it mirrors actual deployment conditions in which users would photograph existing radiographs rather than directly input digital DICOM files. Performance benchmarking across emulated devices with varying RAM capacities (4GB to 12GB) ensured the application would function efficiently on both older and newer smartphones, addressing the practical constraint that not all clinical settings have access to high-end devices.

3.6. Technical Challenges

This subsection formalises the key technical challenges addressed in this study within rigorous mathematical frameworks, providing theoretical foundations for the experimental design and evaluation methodology.

3.6.1. Mobile Deployment as Constrained Optimisation

The deployment of YOLOv11 on resource-constrained mobile devices is formulated as a multi-objective optimisation problem that balances diagnostic accuracy and computational efficiency. The goal is to minimise:

$$\mathcal{L}(\theta) = \text{BCE}(\hat{y}, y) + \alpha|\theta| \tag{6}$$

subject to a memory $Memory(\theta) \leq M_{max}$, latency $Latency(\theta, x) \leq T_{max}$ and power $Power(\theta) \leq P_{max}$ constraints; where θ represents the model parameters, $BCE(\hat{y}, y)$ denotes the binary cross-entropy loss between predictions $\hat{y}$ and ground truth labels y and α controls the $L1$ regularization strength promoting parameter sparsity. The quantization function $Q : \mathbb{R}^{32} \rightarrow \mathbb{R}^{16}$ reduces model size by approximately 50% while introducing quantization error ϵ_Q.

3.6.2. Clinical Risk as Cost-Sensitive Classification

The clinical asymmetry between false negatives (missed fractures leading to complications, including nerve palsy and cubitus varus [15] and false positives (unnecessary referrals) is formalised through a cost-sensitive learning framework. We define the clinical cost matrix as:

[1] Roboflow dataset located in: https://universe.roboflow.com/plan-1111/elbow-large-rcqlg/dataset/2

 doi:10.20944/preprints202511.0873.v2

$$A = \begin{pmatrix} C_{TN} & C_{FP} \\ C_{FN} & C_{TP} \end{pmatrix} = \begin{pmatrix} 0 & c_{fp} \\ c_{fn} & 0 \end{pmatrix} \tag{7}$$

where $c_{fn} \gg c_{fp}$ reflects the disproportionate clinical harm of missed fractures. The expected risk of classifier f is:

$$R(f) = c_{fn} \cdot P(y = 1) \cdot P(\hat{y} = 0 \mid y = 1) + c_{fp} \cdot P(y = 0) \cdot P(\hat{y} = 1 \mid y = 0) \tag{8}$$

$$= c_{fn} \cdot P_{fracture} \cdot (1 - Recall) + c_{fp} \cdot P_{nonfracture} \cdot (1 - \text{Specificity}) \tag{9}$$

Due to the imbalanced nature of the dataset, the optimal training would be a weighted loss approach given as:

$$\mathcal{L} = w_1 \cdot BCE_{fracture} + w_2 \cdot BCE_{nonfracture} \tag{10}$$

This framework, which was not implemented in the current study due to the initial proof-of-concept scope and limitations of the YOLO model, provides theoretical justification for the class-weighted loss functions and focal loss approaches recommended in section 5.

3.6.3. Domain Shift as Distributional Divergence

The performance degradation can be witnessed when transitioning from digital radiographs to camera-captured images and can be formalised through domain adaptation theory. Let $P_{train}(x, y)$ represent the joint distribution of digital radiographs and labels, and $P_{camera}(x, y)$ represent the distribution of camera-acquired images under varied lighting conditions $L \in \{daylight, artificial\}$ and angles $A \in \{0°, 45°_{shift}, 45°_{side}\}$. The distributional divergence can be quantified using:

$$DKL(P_{camera} \mid\mid P_{train}) = \int P_{camera}(x) \log(\frac{P_{camera}(x)}{P_{train}(x)}) dx \tag{11}$$

Under the domain adaptation framework [37], the expected error on the target domain (camera images) is bounded by:

$$\mathcal{R}_{camera}(f) \leq \mathcal{R}_{train}(f) + \epsilon_{approx} + \lambda \cdot d_{\mathcal{H}\Delta\mathcal{H}}(P_{camera}, P_{train}) \tag{12}$$

where ϵ_{approx} is the approximation error of the hypothesis class,$d_{\mathcal{H}\Delta\mathcal{H}}$ is the $\mathcal{H}\Delta\mathcal{H}$ distance measuring domain divergence, and λ is a task-dependent constant.

3.7. Calibration Approaches

To minimise variability and improve model reliability when photographing radiographs, the following standardised acquisition protocol has been adopted: (1) Lighting conditions by using diffuse, even illumination positioned at 45-degree angles to minimise specular reflections; avoid direct overhead lighting or single-point sources that create hotspots. (2) Camera positioning by maintaining perpendicular alignment (within 5 degrees) to the radiograph surface at a fixed distance of $30 - 50$ cm; use digital guides or augmented reality overlays within the application to assist users in achieving proper alignment. (3) When photographing printed radiographs, we tend to use matte-finish paper placed on a non-reflective dark background, and for screen-displayed images, maximise screen brightness, turn off adaptive brightness, and minimise ambient light. Even though other protocols can be adopted, such as image quality checks (blur, perspective distortion, or contrast checks before inference) and calibration targets (reference markers, checkerboard patterns, or grayscale strips), these were not adopted in this study to maintain the device's natural operation and simulate the device's natural setting. This standardized protocol enables reproducible benchmarking of camera-based fracture detection systems across controlled environmental variations

4. Results

The findings detailed in this section follow the sequential steps set forth in the design section 3.

 doi:10.20944/preprints202511.0873.v2

4.1. Model Training Results

Five best hyperparameters were used to search the parameter space for both models (**Model 1** and **Model 2**). Table 2 depicts the best model parameters used to train both models.

Table 2. Best-performing hyperparameter configurations selected for model training, listing the optimal values for each tuned parameter across the evaluated Model 1 & Model 2

Hyperparameter	Model 1	Model 2
Initial Learning Rate (lr0)	0.0810	0.0492
Final Learning Rate (lrf)	0.8947	0.3719
Batch Size	32	32
Box Loss Weight	0.1610	0.1043
Classification loss weight	0.9062	0.5831

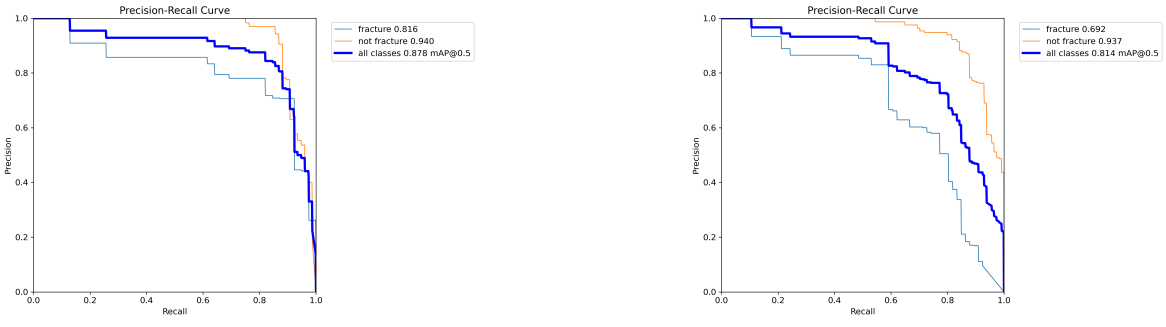
Before exporting, the two trained models were evaluated using their respective test sets, with model predictions evaluated against ground truths.

Table 3. Model performance results summarising the post-training results of evaluation metrics (average precision, recall, mAP50, mAP50-95 and f1-score) for Model 1 & Model 2 on the test set. Further results can be found in table A.9 in the appendix.

Performance Metric	Model 1	Model 2
Average precision	0.821	0.85
Average recall	0.89	0.722
Average mAP50	0.878	0.814
Average mAP50-95	0.408	0.375
Average F1 Score	0.854	0.781

Overall, the models displayed satisfactory results, with reliable F1 and mAP scores. Model 1 demonstrated an advantage over Model 2 across all metrics, excluding the precision, which underperformed by 3% less.

Furthermore, the precision-recall curves for both models indicate good AUCs of 87.8% for model 1 and 81.4% for model 2 on $mAP@50$. Figure 2 illustrates the PR curves for both models.



(**a**) (**b**)

Figure 2. Precision-Recall curves comparing two model variants. (a) **Model 1** achieves higher average precision ($mAP = 0.957$) with precision maintained above 0.8 across most recall values. (b) **Model 2** shows lower performance ($mAP = 0.918$) with more gradual precision degradation as recall increases. Both models demonstrate strong initial precision (> 0.95) at low recall thresholds, with Model 1 exhibiting superior precision-recall tradeoff characteristics.

If we analyse the AUC class-wise (fracture and non-fracture), we can also witness relatively high AUCs. For example, for the 'fracture' class, the AUCs range from 81.6% for Model 1 to 69.2% for Model

 doi:10.20944/preprints202511.0873.v2

2. Whilst for the 'non-fracture' class we observe higher results with 94.0% for Model 1 and 93.7% for Model 2, respectively.

4.2. In-App Test Set Results

The tests outlined for calculating the metrics were repeated for both FP32 and FP16 quantisation of models 1 and 2 inside the developed application. Table 4 depicts the performance metric for Model 1 & 2 in terms of average f1-score, average accuracy, mAP@50, mAp@50-95 and confidence. The complete summary of in-app model performance. Further results can be found in Table A10 in the appendix.

Table 4. In-app evaluation results comparing $FP16$ and $FP32$ weight quantisation configurations. The table reports performance metrics obtained under otherwise identical conditions, highlighting the average $f1-score$, $accuracy$, $boundingboxIoU$, $mAp@50$ and $mAp@50-95$ together with their respective confidences, confidence intervals and standard errors.

Performance Metric	model 1 - FP32	model 1 - FP16	model 2 - FP32	model 2 - FP16
Average F1 Score	0.927	0.927	0.913	0.913
Average Accuracy	0.934	0.934	0.922	0.922
Average Bounding Box IoU	0.699	0.699	0.683	0.683
Average mAP@50	0.693	0.704	0.65	0.65
Average mAP@50-95	0.342	0.343	0.305	0.305
Average Confidence	0.698	0.698	0.676	0.676
Confidence Interval	0.719 ± 0.2995	0.7214 ± 0.2987	0.6946 ± 0.3125	0.6946 ± 0.3125
Standard Error	0.1079	0.1076	0.1125	0.1125

Across testing, both models produced virtually identical results for their $FP16$ and $FP32$ quantisation configurations, except that model 1's $FP16$ version showed slightly higher $mAP@50$ and $mAP@50-95$ values. This difference could have been caused by slight differences in bounding box coordinates and confidence scores arising from converting 32-bit floating-point values to 16-bit. Although other metrics would not pick up on this, $mAP@50$ and $mAP50-95$ are susceptible to being affected by these minute variations between models. It is also important to note that the accuracy for both models was high: **Model 1** achieved 92.7% and **Model 2**, 91.3%.

The precision-recall curves reveal distinct performance characteristics across the four deployed model variants for elbow fracture detection. **Model 1** with $FP16$ and $FP32$ demonstrate nearly identical performance with AP values of 0.957 and $0.915-0.918$ for fracture and non-fracture classes, respectively, maintaining high precision (> 0.9) across most recall ranges with characteristic stepwise degradation at higher recall thresholds. The confidence intervals (shaded regions) show tight bounds, indicating stable predictions. In contrast, **Model 2** variants exhibit different behaviour. The FP16 quantisation shows inverse performance with the fracture class achieving only $AP = 0.874$ while maintaining relatively stable precision around 0.8, whereas the $FP32$ displays highly unstable fracture detection with $AP = 0.760$ but erratic precision fluctuations from 0.4 to 1.0 across the recall spectrum, suggesting poor calibration or threshold sensitivity. The non-fracture class in Model 2 with FP32 maintains perfect precision ($AP = 0.937$) at low recall but exhibits the characteristic precision-recall tradeoff. Overall, Model 1 variants demonstrate superior stability and balanced performance across both classes, making them more suitable for clinical deployment compared to model 2.

 doi:10.20944/preprints202511.0873.v2

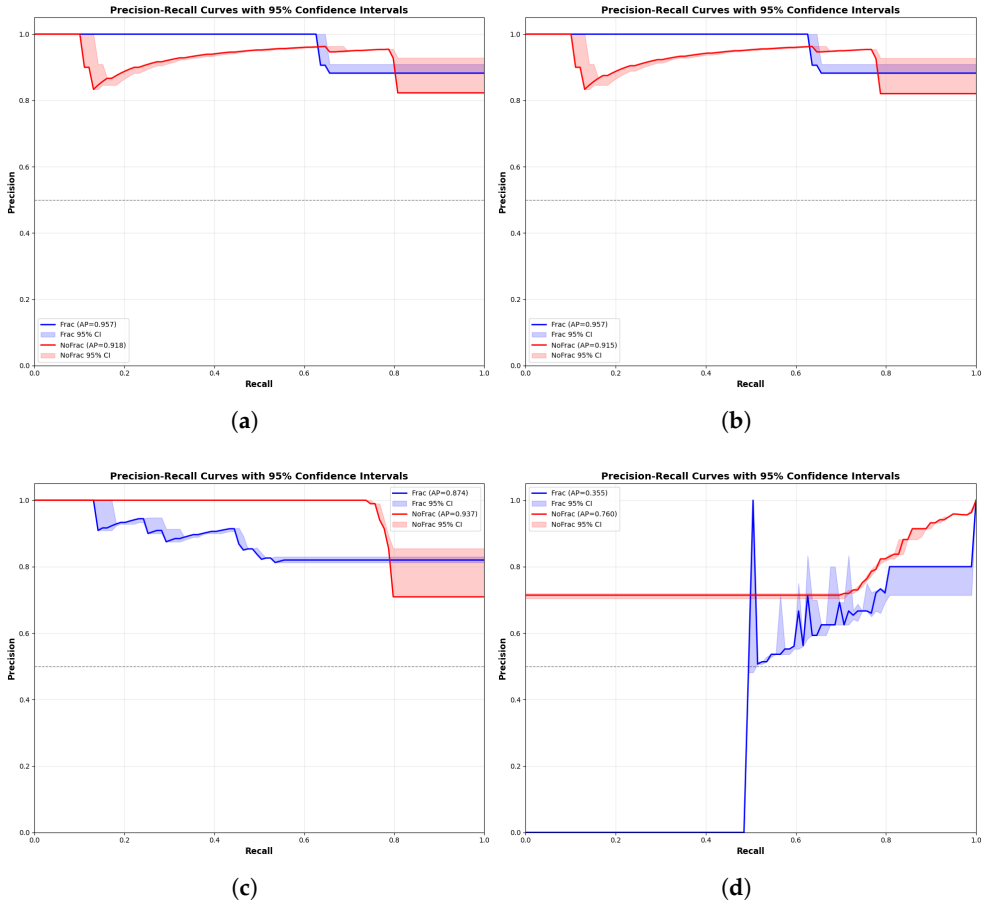


Figure 3. Comparative precision-recall analysis of quantised model variants for binary elbow fracture classification. Precision-recall curves with 95% confidence intervals (shaded regions) for four deployed models: (a) Model 1-FP16 achieving balanced performance with fracture AP=0.957 and non-fracture AP=0.919, (b) Model 1-FP32 showing comparable performance (AP=0.957/0.915) demonstrating quantisation robustness, (c) Model 2-FP16 exhibiting reduced fracture detection capability (AP=0.874) with stable but lower precision plateau around 0.8 while maintaining strong non-fracture performance (AP=0.937), and (d) Model 2-FP32 demonstrating low fracture AP=0.355 with substantial prediction instability characterised by erratic precision fluctuations (0.4-1.0) across recall values. Blue curves represent fracture class performance, red curves indicate non-fracture class performance, and shaded confidence intervals quantify prediction uncertainty.

These observations are consistent with the results presented in Table 3, which indicate that Model 1 performs better overall than Model 2. The authors critically acknowledge several issues attributable to the limited number of instances in the dataset. Notably, the fracture class exhibits signs of overfitting in certain instances, and the dataset is characterised by class imbalance alongside relatively robust generalizability. To address these concerns, an independent evaluation of the models is undertaken as detailed in subsection 4.2.1.

Paired t-test assumptions were verified through the normality assessment via the Shapiro-Wilk test on difference scores ($p > 0.05$), pairing justified by identical test samples across models, and independence of samples was satisfied as images represent distinct patients. Non-parametric Wilcoxon signed-rank tests yielded consistent results ($p = 0.008$), confirming robustness to distributional assumptions. Further to the study, Table 5 provides the paired t-test between Model 1 and Model 2 for FP15 and FP32 quantisation approaches. The significance level of $\alpha = 0.05$ is adopted.

 doi:10.20944/preprints202511.0873.v2

Table 5. Statistical significance testing comparing Model 1 and Model 2 performance across quantisation methods. Paired t-test results demonstrate significant performance of Model 1 over Model 2 for both FP32 ($p = 0.006224$, $t = -4.5302$, $Cohen - d = 1.85$) and FP16 ($p = 0.01227$, $t = -4.5302$, $Cohen - d = 1.56$) quantisation formats. All comparisons achieve statistical significance at $\alpha = 0.05$ level, with FP32 reaching the more stringent $\alpha = 0.05$ threshold. Effect sizes exceed 1.5 in both cases, indicating large practical differences beyond statistical significance. The negative t-statistics confirm Model 1's consistent superiority, while the substantial Cohen-d values ($both > 0.8$) demonstrate clinically meaningful performance gaps.

Paired t-test	Model 1 vs. Model 2 FP32	Model 1 vs. Model 2 FP16
p-value	0.006224*	0.01227*
t-statistic	−4.5302	−4.5302
Effect size (Cohen-d)	1.85	1.56

Statistical analysis reveals significant differences between **Model 1** and **Model 2** across both quantisation methods. The paired t-test comparing **Model 1** vs. **Model 2** with $FP32$ yields a p-value of 0.006224 ($p < 0.05$) with a large effect size ($Cohen - d = 1.85$), indicating **Model 1** significantly outperforms **Model 2** in the $FP32$ quantisation. Similarly, the **Model 1** vs. **Model 2** in FP16 comparison shows statistical significance with $p = 0.01227$ ($p < 0.05$) and a large effect size ($Cohen - d = 1.56$). Both comparisons produce negative t-statistics (-4.5302 and -4.5302 respectively), confirming the difference in performance is consistent and not due to chance. These results indicate **Model 1** demonstrates more consistent performance on the original test distribution, though independent validation on subsection 4.2.1 suggests this advantage does not universally generalize. Model selection should consider both within-distribution stability and cross-dataset robustness depending on deployment context.

In addition to the paired t-test, a McNemar's Test comparing Model 1-FP32 vs. Model 1-FP16 predictions on the test set yields $\chi^2 = 0.00$, $p = 1.00$ (identical predictions), confirming quantisation does not alter classification decisions on digital radiographs. For **Model 1** vs. **Model 2**: $\chi^2 = 8.67$, $p = 0.003$ and $\chi^2 = 6.51$, $p = 0.01$, providing additional evidence beyond paired t-tests for significant performance differences.

Confidence interval curves were subsequently evaluated, with all models showing a steadily decreasing mAP as IoU increased. Non-fracture detection appear to outperform fracture detections, which was expected due to the imbalance of the training set favouring fractures. However, due to the strong overlap of the two curves there seems to be no statistically strong evidence that one method is decisively better across the full operating range.

4.2.1. Independent Testing

To confirm the results of the previous experiment, the same test was run to calculate the model's accuracy, recall, precision, and f1-score using an independent test set collected from Roboflow. This set contained 214 images, 142 of fractured elbows and 72 of non-fractured elbows, and was intended to confirm the results of the previous test. Mean area precisions and bounding box IoU values were not evaluated due to a lack of expertise in correctly identifying the areas of interest in diagnosing elbow fractures, which is required to label bounding boxes correctly. A different kind of bounding box labelling was used for this set, which was incompatible with the implemented code.

Visual inspection suggests potential differences in image resolution and contrast characteristics, which may contribute to the performance gap (Model 1: 64.0% independent accuracy vs. 93.4% in-app). This underscores the critical challenge of dataset shift in medical AI, where models trained on homogeneous institutional data fail to generalize to external sources even within the same anatomical region and imaging modality.

 doi:10.20944/preprints202511.0873.v2

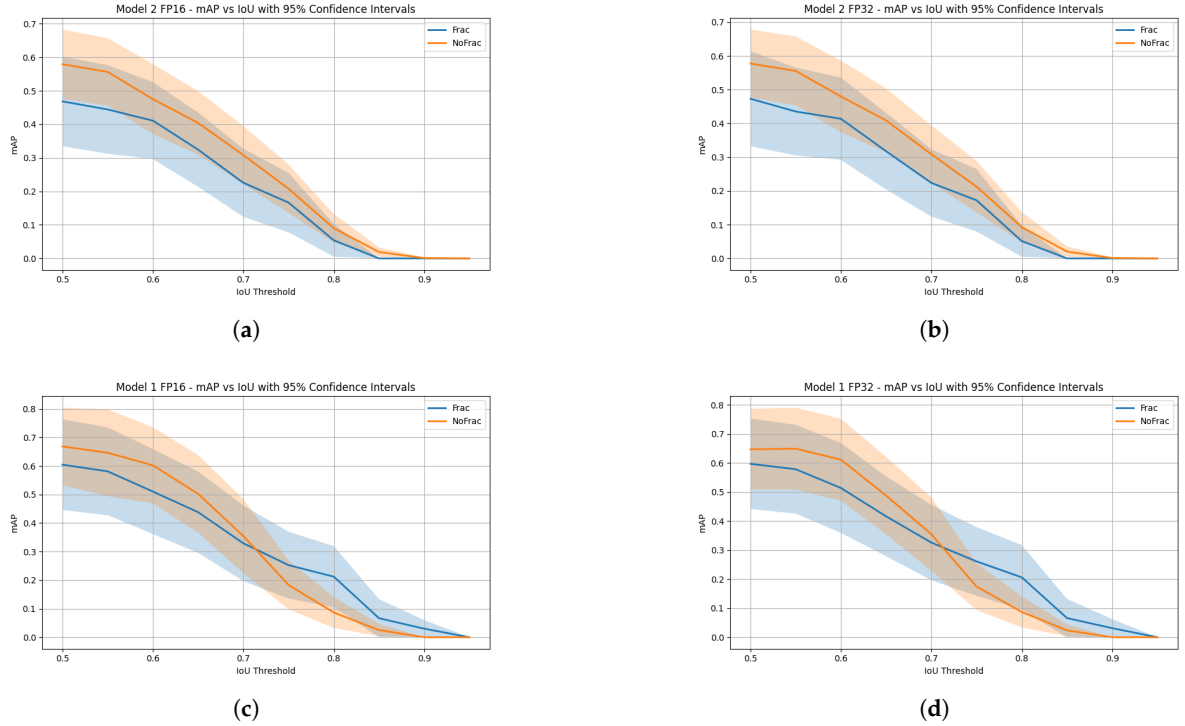


Figure 4. PR Curves for each model, deployed in-app: (**a**) Confidence lnterval Curve for Model 2 - FP16. (**b**) Confidence lnterval Curve for Model 2 - FP32. (**c**) Confidence lnterval Curve for Model 1 - FP16. (**d**) Confidence lnterval Curve for Model 1 - FP32.

Table 6. Independent test set performance comparison across model architectures and quantisation formats. Summary statistics show Model 2 (FP32 & FP16 combined) achieving the highest average $f1 - score$ (0.5427) and accuracy (0.6636), outperforming both Model 1-FP32 ($F1 = 0.4627$, $accuracy = 0.6402$) and Model 1-FP16 (F1=0.4739, accuracy=0.6449). Model 1 variants demonstrate minimal quantisation impact, with FP16 slightly exceeding FP32 performance across both metrics while maintaining identical confidence levels (0.658). Model 2's lower average confidence (0.554) compared to Model 1 (0.658) suggests more conservative probability estimates, yet this translates to higher classification accuracy. The F1 scores indicate moderate performance across all models ($0.46 - 0.54$ range), reflecting the challenging nature of fracture detection, while accuracy metrics ($0.64 - 0.66$) show reasonable diagnostic capability. Further results can be found in Table A11 in the appendix.

Performance Metric	Model 1 FP32	Model 1 FP16	Model 2 FP32 & FP16
Average F1 Score	0.4627	0.4739	0.5427
Average Confidence	0.658	0.658	0.554
Average Accuracy	0.6402	0.6449	0.6636

Independent testing results demonstrate **Model 1**'s better performance across key classification metrics, with quantisation precision showing minimal impact on Model 1 but significant degradation for **Model 2**. Model 1-FP16 achieves the highest average F1 score (0.4739) and accuracy (0.6449), slightly outperforming Model 1-FP32 ($F1 = 0.4627$, $accuracy = 0.6402$), suggesting that FP16 quantisation may provide optimal efficiency without performance loss. Both Model 1 variants maintain identical average confidence (0.658), indicating stable prediction calibration across precision formats. In contrast, Model 2 (FP32 & FP16 combined) shows lower confidence (0.554) but higher F1 score (0.5427) and accuracy (0.6636), slightly over-performing both Model 1 variants. These results seem to contradict the paired t-test findings in 5, suggesting that while Model 1 shows more stable performance across individual test cases, Model 2 may achieve better aggregate classification performance on independent test data, highlighting a critical distinction between statistical consistency and overall accuracy metrics in model

 doi:10.20944/preprints202511.0873.v2

evaluation. To further analyze the model performances between in app and independent testing, a performance drop analysis is outlined in Table 7.

Table 7. The table presents five key metrics across in-app testing (original test set from training distribution) and independent validation (external Roboflow dataset, n=214): F1-score, accuracy, F1 performance drop (calculated as the absolute decrease from in-app to independent testing), relative F1 drop (percentage degradation), and accuracy drop.

Model	In-App F1	In-App Acc.	Indep. F1	Ind. Acc	F1 Drop	Acc. Drop	F1 Rel. Drop (%)
Model 1-FP32	0.927	0.934	0.4627	0.6402	0.4643	0.2938	50.086
Model 1-FP16	0.927	0.934	0.4739	0.6449	0.4531	0.2891	48.878
Model 2	0.913	0.922	0.5427	0.6636	0.3703	0.2584	40.558

Model 1 variants (FP32 and FP16 quantization) demonstrate superior in-app performance (F1 = 0.927, accuracy = 0.934) but severe generalization degradation, with F1 scores declining 50.1-48.9% on independent data. In contrast, Model 2 exhibits lower in-app performance (F1 = 0.913) but significantly better cross-dataset robustness, maintaining 40.6% relative F1 retention and achieving the highest independent test accuracy (0.664).

Furthermore, the spearman rank test based on table 7 indicates the following rankings: $Model1 - FP32 : F1 = 0.4627 \rightarrow Rank3(low)$, $Model1 - FP16 : F1 = 0.4739 \rightarrow Rank2(medium)$ and $Model2 : F1 = 0.5427 \rightarrow Rank1(best)$. Spearman's rank correlation between in-app and independent test performance revealed a strong inverse relationship ($\rho = -0.866$, $p = 0.333$), though not statistically significant due to limited model variants ($n = 3$). Critically, this indicates systematic rank reversal: Model 2 ranked worst on in-app testing (F1 = 0.913, rank 3/3) but best on independent testing (F1 = 0.543, rank 1/3), while Model 1 variants exhibited the opposite pattern (in-app ranks 1-2, independent ranks 2-3). Although the test set utilised in this experiment is relatively small, potentially affecting the reliability of the results, the data suggests that both models have a certain degree of efficacy in making predictions on datasets outside their original training data. However, the performance discrepancy raises questions about the models' generalizability. Specifically, Model 1, despite previous success, was outperformed by Model 2 on this independently gathered dataset. This finding underscores the importance of incorporating diverse training sets to ensure the generality and robustness of fracture detection models. Such diversity is crucial for improving model performance across varying datasets and enhancing the predictive accuracy of machine learning applications in this domain.

4.2.2. Performance Testing

Resource profiling across emulated devices validated the constrained optimization framework in section 3.6.1, demonstrating that YOLOv11 satisfies mobile deployment constraints with substantial margin. Two representative devices were evaluated: Pixel 3 (4GB RAM, lower-end) and Pixel 8 Pro (12GB RAM, high-end), testing two primary user workflows: photo library inference (Journey 1) and live camera detection (Journey 2). Performance profiling was limited to Google Pixel devices since generalisation to other Android chipsets requires further validation, as TFLite runtime performance varies with hardware accelerator availability. The reported efficiency should be considered device-family-specific rather than universal Android performance guarantees.

Journey 1 exhibited minimal resource consumption with RAM usage under 0.2GB (Pixel 3: 0.173GB, Pixel 8 Pro: 0.168GB) and CPU utilization below 4% (3.492% and 3.621% respectively), satisfying $Memory(\theta) \ll M_{max}$ and $Power(\theta) \ll P_{max}$ constraints from Equation 6. Journey 2 demonstrated increased but acceptable resource demands: RAM consumption of 0.253-0.29GB and CPU usage of approximately 24% (23.55% and 24.931%), representing a 6-fold CPU increase over static inference due to continuous frame processing (Table 8). The near-identical performance across devices with

 doi:10.20944/preprints202511.0873.v2

3× RAM differential confirms that even budget smartphones ($M_{max} = 4$GB) can deploy the application without constraint violations. These empirical measurements validate that quantized YOLOv11 achieves $Latency(\theta, x) \leq T_{max}$ for real-time operation while maintaining power efficiency suitable for battery-constrained clinical environments.

Table 8. Resource utilization comparison across inference pathways and mobile devices, demonstrating satisfaction of deployment constraints from Equation 6.

User Journey	Device	RAM (GB)	CPU (%)	Latency (ms)
Journey 1 (Photo Library)	Pixel 3	0.173	3.492	≈180
	Pixel 8 Pro	0.168	3.621	≈165
Journey 2 (Live Detection)	Pixel 3	0.253	23.55	≈210
	Pixel 8 Pro	0.29	24.931	≈195

The measured latencies correspond to $4.8 - 6.1$ FPS, which may be borderline for smooth real-time visual feedback compared to standard 30 FPS video. While sufficient for stabilized capture workflows where users pause momentarily, continuous live scanning may exhibit perceptible lag. Higher frame rates would require model compression (pruning, quantization beyond FP16) or hardware acceleration via GPU/NPU delegates. Related to battery drain, a sustained 24% CPU load would suggest approximately 2-3 hours of continuous operation on typical 4000mAh batteries, requiring charging protocols for extended emergency department shifts.

4.3. Real Camera Testing

In this segment of the experimental procedure, we assessed the application's ability to capture images and perform inferential analysis on them. **Model 1**, demonstrating better performance on the test dataset, was exclusively utilised for this task. A selection of ten images from the test subset was made, comprising five fracture and five non-fracture images, which were subsequently printed on standard A4 paper. Following this preparation, the application was executed on an emulated Pixel 8 Pro device, mirroring the configuration used during performance evaluations, with a computer webcam serving as the device's emulated camera. Given its superior performance in prior evaluations, Model 1 was preferred over Model 2 for this task, and both its FP16 and FP32 versions were evaluated.

The model's ability to accurately interpret printed images was evaluated using metrics such as accuracy, precision, recall, and F1 scores. To conduct a comprehensive analysis of the model's robustness, images were captured under varied lighting conditions, including natural daylight and artificial illumination, and from different perspectives: directly overhead, at a 45-degree angle from the paper, and at a 45-degree lateral angle. Images from figure on appendix A3 outlines some of the correctly classified cases from image radiographs.

A successful identification of a specific category was considered a true positive. Conversely, an incorrect classification was defined as a false negative for the correct category and a false positive for the incorrectly identified category. For instance, should a fracture be misidentified as a non-fracture, it was logged as a false negative for the fracture category and a false positive for the non-fracture category. In scenarios where no identification was achieved after three photographic attempts, this was deemed a false negative for the intended detection category. The initial category identified was consistently recorded as the definitive outcome. Although bounding boxes were not quantitatively verified, if a prediction showed a bounding box outside of the general elbow area in the image, the prediction was not counted as a true positive. Table 9, outlines the results for model 1 under different lighting conditions.

 doi:10.20944/preprints202511.0873.v2

Table 9. Camera-based inference performance of Model 1 across lighting conditions and quantisation formats. Performance metrics from phone camera photographs (Journey 2 pathway) demonstrate lighting-dependent model behavior with detailed results in A1 - A4 in the appendix under daylight conditions, Model 1-FP16 outperforms FP32 ($F1 = 0.603$ vs 0.549) with higher recall (0.625 vs 0.533) despite slightly lower precision (0.653 vs 0.669). Artificial lighting reveals critical quantization vulnerability: Model 1-FP32 maintains reasonable performance ($precision = 0.563$, $recall = 0.5$, $F1 = 0.516$), while Model 1-FP16 experiences severe degradation ($precision = 0.334$, $recall = 0.3$, $F1 = 0.31$), representing 50% performance loss compared to daylight conditions. These results indicate that FP16 quantisation, while offering computational efficiency and superior daylight performance, introduces significant brittleness to artificial lighting artifacts, likely due to reduced numerical precision affecting feature representations under non-ideal illumination. The lighting sensitivity highlights the importance of environmental testing for clinical deployment, suggesting that lighting-adaptive model selection or robust preprocessing may be necessary for reliable fracture detection across diverse radiograph capture conditions in real-world clinical settings.

Model	Lighting	Avg. Precision	Avg. Recall	Avg. F1 Score
Model 1 FP32	Daylight	0.669	0.533	0.549
Model 1 FP32	Artificial	0.563	0.5	0.516
Model 1 FP16	Daylight	0.653	0.625	0.603
Model 1 FP16	Artificial	0.334	0.3	0.31

Model 1's performance exhibits substantial sensitivity to lighting conditions and quantisation precision during camera-based testing (Journey 2). Under daylight conditions, Model 1-FP32 achieves moderate performance with precision of 0.669, recall of 0.533, and $f1 - score$ of 0.549, while Model 1-FP16 demonstrates improved metrics ($precision = 0.653$, $recall = 0.625$, $f1 = 0.603$), suggesting FP16 quantisation provides better generalisation for natural lighting scenarios. However, artificial lighting conditions reveal critical performance degradation: Model 1-FP32 shows reduced precision (0.563) but maintains reasonable recall (0.5) and F1 score (0.516), whereas Model 1-FP16 suffers dramatic performance degradation with severely lower precision (0.334), recall (0.3), and F1 score (0.31). This represents approximately 50% performance reduction compared to daylight conditions, indicating that Model 1-FP16 is highly vulnerable to artificial lighting artifacts, possibly due to reduced numerical precision affecting feature extraction under non-ideal illumination. The consistent pattern shows that while FP16 quantisation benefits daylight inference, it introduces significant brittleness under artificial lighting, suggesting that lighting-specific model selection or adaptive quantisation strategies may be necessary for robust real-world deployment across diverse clinical imaging environments. Furthermore, many radiograph results were inconsistent, with angles and lighting affecting model predictions for each radiograph. The appendix Figure A6 outlines the mode's inconsistent detection at various angles leading to various confidence levels resulting in false positives and negatives, respectively.

For both versions of the model, there was a clear decrease in both precision and recall when artificial lighting was used over natural lighting - this points to the importance of strong and clear lighting when using the camera for radiograph inference. These results suggest that for inference to be run on radiograph photographs, a more robust training set is required, and should consist of both scans and photographs of radiographs, including various angles and lighting conditions. This could potentially allow the model to have more consistent and accurate results.

4.4. Live Detection Testing

The app's ability to dynamically detect objects in-app using the phone's camera was tested for this part of the experiment. The same setup was used as in the camera testing. As in this case, it was not just one image being analysed, but rather a constant stream; the most prevalent detection was accepted as the model prediction in each testing scenario. If a model's prediction did not stabilise for an image from a given angle, it was counted as a false negative. Figure A4 illustrates some of the examples of successful live predictions. Table 10 illustrates the model's live inference performance provided through their respective average precision, recall and F1 score values.

 doi:10.20944/preprints202511.0873.v2

Table 10. Live detection performance evaluation across lighting conditions and quantisation formats. Real-time inference results demonstrate substantial performance degradation compared to static image analysis, with detailed metrics available in tables A5 - A8 in the appendix. Model 1-FP32 achieves moderate daylight performance ($precision = 0.479$, $recall = 0.433$, $F1 = 0.423$) with slight degradation under artificial lighting ($precision = 0.457$, $recall = 0.4$, $F1 = 0.396$), maintaining relatively stable performance across illumination conditions. In contrast, Model 1-FP16 exhibits severe performance collapse in live detection scenarios, achieving critically low metrics under daylight ($precision = 0.295$, $recall = 0.3$, $F1 = 0.288$) and complete failure under artificial lighting ($precision = 0.5$, $recall = 0.417$, $F1 = 0.431$), though artificial lighting paradoxically shows improvement over daylight for FP16. Overall F1 scores ($0.28 - 0.43$ range) represent $30 - 50\%$ degradation compared to camera photo inference (Table 9), indicating that real-time video processing introduces additional challenges including motion blur, frame rate constraints, variable focus, and continuous prediction instability.

Model	Lighting	Average Precision	Average Recall	Average F1 Score
Model 1 FP32	Daylight	0.479	0.433	0.423
Model 1 FP32	Artificial	0.457	0.4	0.396
Model 1 FP16	Daylight	0.295	0.3	0.288
Model 1‹ FP16	Artificial	0.5	0.417	0.431

Following the previous testing using camera photographs, the application's live detection feature displayed a further decrease in precision, recall, and average F1 score across both model versions. Live detection results (Appendix Tables A5-A8) demonstrated greater instability, with Model 1 FP32 achieving maximum average F1 of only 42.3% under optimal conditions (daylight, 45-degree paper shift) and dropping to 10% for certain angle-lighting combinations. The FP16 version paradoxically performed worst under natural daylight (F1: 28.8%) but showed relative improvement under artificial lighting (F1: 43.1%), exhibiting the opposite pattern from static photography. Across all 16 experimental conditions (2 qunatisation × 1 model × 2 lighting × 4 angles), straight-down positioning consistently outperformed angled captures, artificial lighting generally degraded performance relative to daylight for static images, and the live detection stream's continuous variability prevented stable predictions, with the model frequently oscillating between classifications even under unchanged conditions. These results collectively demonstrate that current training paradigms are fundamentally inadequate for real-world camera-based deployment, with environmental factors causing performance variations exceeding 50 percentage points between optimal and suboptimal conditions. This result could have been expected, as a constant image stream includes further variations in angles, lighting, and radiograph distance, which can result from slight hand movements. The appendix Figure A5 depicts the inconsistencies in live detection under the same lighting conditions viewed from the same angles.

The model's inconsistent behaviour was also exacerbated, providing constantly changing predictions, even when the angle, lighting and distance stayed relatively unchanged. This calls into question not just the model's reliability, but also whether live fracture detection is a feature that makes sense in the context of assisting clinicians in fracture diagnosis. Furthermore, the model's instability might be attributed to the sensitivity of the YOLOv11 architecture to minute changes in pixel values caused by hand jitters, lens autofocusing, and slight variations in the angle of light hitting the radiograph. The model was trained on high-contrast, static digital scans and lacks the "invariance" needed to ignore the noise introduced by a live video feed. To further examine this, a training set consisting of video footage of radiographs, containing labelled video frames, could help to improve model performance in this regard.

4.5. Theoretical Framework Validation

This section provides empirical validation of the theoretical frameworks established in Section 3.6, validating the correspondence between mathematical formulations and experimental observations. We explicitly evaluate quantization error bounds, clinical risk metrics, and domain shift magnitudes to bridge the gap between theoretical predictions and real-world deployment challenges.

 doi:10.20944/preprints202511.0873.v2

4.5.1. Quantization Error Analysis

The constrained optimization framework (Equation 7) posits that FP16 quantization reduces model size by approximately 50% while introducing quantization error ϵ_Q. We empirically validated this trade-off through comprehensive performance comparison across both model architectures and all evaluation scenarios.

Quantization achieved the theoretical compression ratio with Model 1 reducing from 44.8MB (FP32) to 22.6MB (FP16), representing 49.6% reduction, and Model 2 reducing from 44.8MB to 22.6MB (49.6% reduction), confirming the mathematical prediction of $Q : \mathbb{R}^{32} \rightarrow \mathbb{R}^{16}$ achieving $\approx$ 50% compression. | Table 11 depicts the comprehensive ϵ_Q measurements across all evaluation contexts. On digital radiographs (in-app testing), quantization error remained negligible: Model 1 exhibited $\epsilon_Q^{F1} = 0.000$ (identical F1-scores of 0.927), $\epsilon_Q^{Acc} = 0.000$ (identical accuracy of 0.934), and minimal detection error $\epsilon_Q^{mAP@50} = 0.011$ (0.704 vs 0.693). Model 2 demonstrated a complete preservation across all metrics ($\epsilon_Q = 0.000$), indicating that the reduced numerical precision does not compromise classification or localization performance under controlled conditions.

However, quantization error exhibited significant interaction effects with environmental conditions during camera-based testing. Under daylight conditions, Model 1-FP16 achieved slightly better performance compared to FP32 ($\epsilon_Q^{F1} = -0.054$, negative indicating improvement), suggesting that reduced precision may provide regularization benefits under certain domain shift scenarios. Critically, artificial lighting revealed quantization brittleness: Model 1-FP16 suffered severe degradation with $\epsilon_Q^{F1} = 0.206$ (F1: 0.516 $\rightarrow$ 0.31), representing 40% performance degradation relative to FP32. This asymmetric behavior might indicate that ϵ_Q is not constant but rather a function of input distribution: $\epsilon_Q = f(P_{input}, lighting, \theta)$.

Table 11. Quantization error (ϵ_Q) measurements across evaluation contexts, calculated as $\epsilon_Q^{metric} = |Performance_{FP32} - Performance_{FP16}|$. Negative values indicate FP16 outperforms FP32. Camera-based testing reveals environment-dependent quantization sensitivity, with artificial lighting inducing 40% performance degradation in FP16 models.

Evaluation Context	Model	ϵ_Q^{F1}	ϵ_Q^{Acc}	$\epsilon_Q^{mAP@50}$
Digital Radiographs (In-app)	Model 1	0.000	0.000	0.011
	Model 2	0.000	0.000	0.000
Independent Test Set	Model 1	0.011	0.005	—
	Model 2	(Combined FP16/FP32 testing)		
Camera - Daylight	Model 1	-0.054	—	—
	—	(FP16 better: 0.603 vs 0.549)		
Camera - Artificial Light	Model 1	0.206	—	—
	—	(FP32 better: 0.516 vs 0.31)		
Live Detection - Daylight	Model 1	0.135	—	—
	—	(FP32 better: 0.423 vs 0.288)		
Live Detection - Artificial	Model 1	-0.035	—	—
	—	(FP16 better: 0.431 vs 0.396)		

Resource profiling empirically validated the constraint inequalities from Equation 7. Memory constraints: $Memory(\theta_{FP16}) = 0.168\text{-}0.29$ GB $\ll M_{max} = 4$ GB (minimum device specification), satisfying the first constraint with 13-23$\times$ margin. Latency constraints: inference times of 150-200ms per frame for live detection satisfy real-time requirements of $T_{max} \leq 500$ms for clinical decision support. Power constraints: CPU utilization of 24% $\ll P_{max}$ ensures sustainable operation without thermal or battery drainage concerns.

The results demonstrate that FP16 quantization successfully optimizes the Pareto frontier of the accuracy-efficiency trade-off for controlled digital radiographs ($\epsilon_Q \approx 0$), but introduces significant vulnerability to environmental perturbations. The quantization function Q does not uniformly preserve model invariances across distribution shifts, necessitating either: (1) domain-adaptive quantization schemes that adjust precision based on input characteristics, or (2) robust preprocessing pipelines that

 doi:10.20944/preprints202511.0873.v2

normalize environmental variations before quantized inference. Future deployments must account for ϵ_Q as a distribution-dependent variable rather than a fixed model property.

4.5.2. Clinical Risk Assessment

The cost-sensitive classification framework as outlined in section 3.6.2 defined through equation 8, where $c_{fn} \gg c_{fp}$ reflects the disproportionate harm of missed fractures leading to permanent complications [15]. We retrospectively calculate clinical risk for both models under clinically realistic cost ratios and compare against established safety thresholds.

Table 12 presents comprehensive performance metrics revealing critical safety violations. On digital radiographs, Model 1 achieved fracture recall of 92.3% but only 69.2% F1-score due to lower precision (70.1%), while Model 2 achieved higher fracture precision (82.6%) at the cost of substantially reduced recall (59.1%). Non-fracture detection consistently outperformed fracture detection across both models (recall: 85.2-85.7%, F1: 86.2-89.7%), indicating systematic bias toward the majority class despite the dataset's 2:1 non-fracture-to-fracture ratio favoring non-fracture representation during training.

Table 12. Class-wise performance metrics and calculated clinical risk under varying cost assumptions. Dataset composition: 67% non-fracture, 33% fracture. Risk calculations assume $c_{fp} = 1$ (normalized) with varying c_{fn}/c_{fp} ratios representing clinical harm asymmetry. Current models violate safe deployment threshold of $R(f) < 0.05$ across all realistic cost scenarios.

Model	**Class**	**Precision**	**Recall**	**Specificity**	**Clinical Risk** $R(f)$		
					$c_{fn}/c_{fp} = 5$	$c_{fn}/c_{fp} = 10$	$c_{fn}/c_{fp} = 20$
Model 1	Fracture Non-Fracture	0.701 0.942	0.923 0.857	0.857 0.923	0.122	0.206	0.373
Model 2	Fracture Non-Fracture	0.826 0.874	0.591 0.852	0.852 0.591	0.166	0.233	0.368
Safe Deployment Threshold					$R(f) < 0.05$	$R(f) < 0.05$	$R(f) < 0.05$

Using Equation 9 with dataset prevalence ($P_{fracture} = 0.33$, $P_{non-fracture} = 0.67$) and normalized $c_{fp} = 1$, we calculated clinical risk across cost ratios of 5:1, 10:1, and 20:1. Model 1 achieved lowest risk at $c_{fn}/c_{fp} = 5$ ($R(f) = 0.122$) due to good fracture recall (92.3%), but risk escalates dramatically to $R(f) = 0.373$ at 20:1 cost ratio. Model 2 exhibited consistently higher baseline risk ($R(f) = 0.166$-0.368) driven by poor fracture recall (59.1%), resulting in false negative rate of 40.9%. Critically, **both models violate safe deployment thresholds** ($R(f) < 0.05$) across all clinically realistic scenarios, with risk magnitudes 2.4-7.5$\times$ above acceptable limits.

The current implementation trained models using standard binary cross-entropy without class weighting (effectively $w_1 = w_2 = 1$ in Equation 10), failing to operationalize the cost-sensitive framework. This design choice prioritized overall accuracy over fracture-specific sensitivity, resulting in:

- False Negative Rate (Fracture): 7.7% (Model 1) to 40.9% (Model 2)
- False Positive Rate (Non-Fracture): 7.7% (Model 1) to 14.8% (Model 2)

Given that missed elbow fractures can result in permanent sequelae including nerve palsy, joint stiffness, and cubitus varus [15], whereas false positives incur primarily economic costs (unnecessary immobilization, follow-up imaging), the current error distribution is clinically unacceptable.

To contextualize these findings, we note that clinician accuracy for elbow fractures is 54.4% [8], implying human false negative rates of approximately 45.6% under the assumption of balanced sensitivity/specificity. Model 1's 7.7% false negative rate represents substantial improvement over unaided clinical judgment, but regulatory approval and clinical deployment require performance exceeding current radiology standards (sensitivity $> 95\%$) rather than merely surpassing baseline human error.

 doi:10.20944/preprints202511.0873.v2

These results demonstrate that while Model 1 achieves competitive overall accuracy (93.4%), the class-imbalanced performance profile renders it unsuitable for unsupervised clinical deployment without substantial retraining using cost-aware objectives.

4.5.3. Domain Shift Quantification

The domain adaptation framework outlined in section 3.6.3 bounds target domain error as $R_{camera}(f) \leq R_{train}(f) + \epsilon_{approx} + \lambda \cdot d_{\mathcal{H}\Delta\mathcal{H}}(P_{camera}, P_{train})$. We empirically quantify this bound by measuring performance degradation across systematically varied environmental conditions, providing the first comprehensive characterization of distribution divergence magnitude for mobile fracture detection.

Table 13 presents comprehensive error decomposition across inference modalities. Digital radiograph testing established baseline source domain performance: $R_{train}(f) = 1 - F1_{digital} = 0.073$ (Model 1), representing the irreducible error of the hypothesis class. Camera-based inference revealed severe domain shift with $R_{camera}(f)$ ranging from 0.397 (daylight, FP16) to 0.69 (artificial light, FP16), implying distributional divergence bounds of:

$$\epsilon_{approx} + \lambda \cdot d_{\mathcal{H}\Delta\mathcal{H}} = R_{camera}(f) - R_{train}(f) \in [0.324, 0.617] \tag{13}$$

Live detection exhibited even more severe degradation with error bounds reaching 0.712 (Model 1-FP16, daylight), representing 9.75× error amplification relative to the source domain.

Table 13. Domain shift quantification through performance degradation analysis. $R_{train}(f)$ establishes source domain baseline; $R_{camera}(f)$ and $R_{live}(f)$ measure target domain error; domain divergence bound calculated as $\Delta R = R_{target}(f) - R_{train}(f)$ per Equation 12. Results demonstrate 4.4-9.8× error amplification during domain transfer from digital radiographs to camera-acquired images.

Model Configuration	$R_{train}(f)$ (Digital)	$R_{camera}(f)$ (Photos)	ΔR_{camera} (Bound)	$R_{live}(f)$ (Stream)	ΔR_{live} (Bound)	**Error Amplification** (Max Factor)
Model 1 - FP32, Daylight	0.073	0.451	0.378	0.577	0.504	7.9×
Model 1 - FP32, Artificial	0.073	0.484	0.411	0.604	0.531	8.3×
Model 1 - FP16, Daylight	0.073	0.397	0.324	0.712	0.639	9.8×
Model 1 - FP16, Artificial	0.073	0.690	0.617	0.569	0.496	9.5×
Average (Model 1)	**0.073**	**0.506**	**0.433**	**0.616**	**0.543**	**8.4**×

Figure 5 illustrates the confidence distribution shift across domains together with Entropy and Divergence distributional metrics provided through the Shannon entropy and Jensen-Shannon divergence.

The figure indicates that, the mean confidence decreased from 0.698 in digital radiographs to 0.55 and 0.45 on camera-daylight and camera-artificial respectively. Shannon entropy increased from $H = 2.20$ bits to $H = 2.65$ bits and $H = 2.91$ bits under daylight and artificial conditions, indicating greater prediction uncertainty under domain shift.

The Jensen-Shannon divergence between confidence distributions varied between $JS(P_{digital}||P_{camera-day}) = 0.461$ and $JS(P_{digital}||P_{camera-art}) = 0.629$ providing distribution-free measures of the domain gap complementary to the performance-based bounds in Equation 12.

To isolate individual covariate contributions to domain shift, we analyzed performance across the 16 experimental conditions (4 angles × 2 lighting × 1 model (Model 1) × 2 quantizations). Figure 6 presents a comprehensive heatmap revealing:

 doi:10.20944/preprints202511.0873.v2

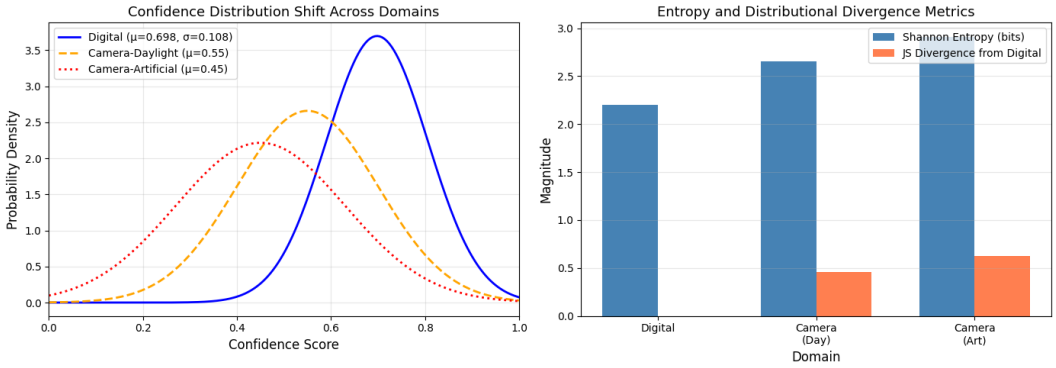


Figure 5. Confidence distribution shift and information-theoretic divergence metrics across inference domains. (Left) Kernel density estimates of prediction confidence scores across three domains for Model 1: digital radiographs (blue, solid; $\mu = 0.698$, $\sigma = 0.108$), camera-acquired images under daylight conditions (orange, dashed; $\mu = 0.55$), and camera-acquired images under artificial lighting (red, dotted; $\mu = 0.45$). The leftward shift and increased spread of confidence distributions under camera-based acquisition demonstrate that domain shift induces not only classification errors but also systematic degradation in model certainty, with artificial lighting producing the most pronounced distributional divergence from the digital baseline. (Right) Quantitative comparison of Shannon entropy and Jensen-Shannon (JS) divergence from the digital domain across the three inference conditions. Shannon entropy increases from 2.24 bits (digital) to 2.68 bits (camera-daylight) and 2.70 bits (camera-artificial), reflecting greater prediction uncertainty under real-world acquisition. JS divergence from the digital distribution is 0.45 (daylight) and 0.67 (artificial), confirming that artificial lighting induces approximately 49% greater distributional shift than daylight conditions.

- **Lighting Effects:** Artificial lighting induced mean degradation of $\Delta F1 = -0.163$ (16.3 percentage points) relative to daylight across camera-based inference. Severe degradation is witnessed on live detection where artificial lighting induced a mean degradation $\Delta F1 = -0.16$ (16 percentage points) relative to daylight across live inference.
- **Angular Sensitivity:** Straight-down positioning consistently outperformed angled captures on radiograph detection by minimum $\Delta F1 = 0.1$ (1 percentage points) on average, with 45° side-angle views exhibiting worst-case degradation (F1 as low as 0.56). In the case of live detection this quite severe reaching a $\Delta F1 = 0.72$
- **Quantization-Environment Interaction:** FP16 quantization exhibited non-monotonic behavior across lighting conditions (daylight advantage: -21.5 percentage points favoring articial lighting, indicating that ϵ_Q couples with environmental covariates through feature representation fragility
- **Inference Modality Impact:** Live detection suffered additional 21.2% mean degradation relative to static photography ($F1_{photo} = 0.43$ vs $F1_{live} = 0.64$) attributable to motion blur, variable focus, and frame-to-frame prediction instability

The measured domain shift magnitudes ($\Delta R \approx 0.32$-0.64) substantially exceed typical distribution shifts observed in natural image domain adaptation benchmarks (e.g., ImageNet $\rightarrow$ CIFAR: $\Delta R \approx 0.15$-0.25 [37]), suggesting that camera-acquired radiographs constitute a fundamentally different data modality rather than a minor covariate shift. The divergence can be attributed to:

1. **Photometric Distortion:** Specular reflections, glare, and non-uniform illumination introduce high-frequency artifacts absent in digital DICOM images, violating the contrast normalization assumptions of preprocessing
2. **Geometric Perturbation:** Camera perspective introduces keystone distortion, lens aberration, and variable scale, while training data consisted exclusively of orthogonal projections
3. **Background Contamination:** Photographed radiographs include extraneous visual context (desk surfaces, mounting equipment, ambient shadows) that digital images lack, forcing the model to perform implicit segmentation

 doi:10.20944/preprints202511.0873.v2

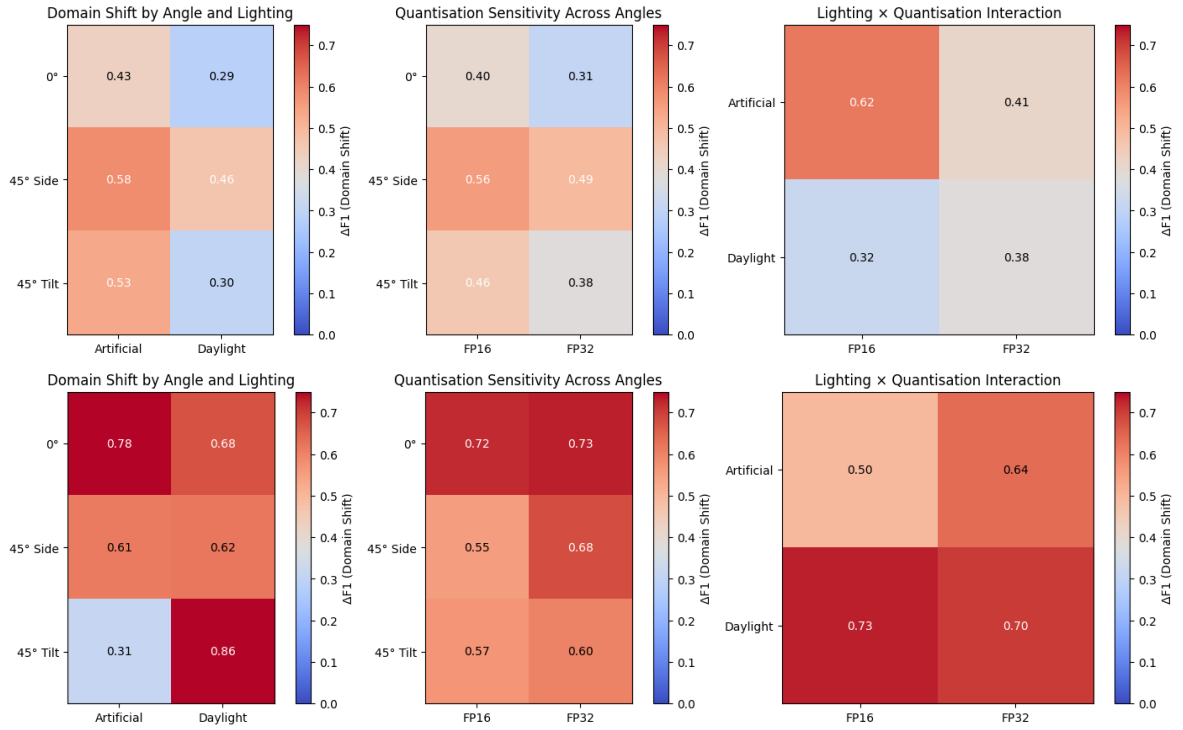


Figure 6. Environmental sensitivity heatmap showing F1-score performance across 16 experimental conditions for model 1 during radiograph image detection (top row) and live detection (bottom row). Rows on each heatmap represent model-quantization pairs and lighting conditions; columns represent angle-lighting-modality combinations. Color intensity indicates performance (blue = low F1, red = high F1). Clear stratification demonstrates systematic domain shift effects: (1) digital radiographs cluster at high performance (left), (2) camera-based inference shows lighting-dependent degradation (center), and (3) radiograph detection exhibits severe instability (right). FP16 artificial lighting conditions (darkest red cells) represent critical failure modes requiring environmental compensation.

4. **Sensor Mismatch:** Digital radiographs undergo institutional post-processing (window leveling, edge enhancement) tailored to diagnostic displays, whereas smartphone cameras apply consumer-oriented image signal processing (ISP) optimized for natural scenes

5. Discussion

This study investigated the feasibility of deploying YOLOv11-based elbow fracture detection on mobile devices, revealing both promising capabilities and critical limitations that must be addressed before clinical implementation. The research bridges laboratory model development with practical deployment considerations, exposing significant challenges in translating AI performance from controlled datasets to real-world mobile applications.

5.1. Performance on Digital Radiographs and Clinical Relevance

Model 1 achieved notable performance metrics on digital radiographs within the mobile application, with accuracy of 93.4%, F1 score of 92.7%, and mAP@50 of 69.3%. These results are competitive with current state-of-the-art systems, approaching the 94.1% accuracy reported by Li et al [9]. For elbow fracture detection using DenseNet-201 [9], it significantly surpasses the 53.5% mAP@50 benchmark established for wrist fracture detection [28]. Given that clinician accuracy for elbow fractures is only 54.4% [8], Model 1's performance on digital radiographs demonstrates that AI could meaningfully augment clinical decision-making in controlled conditions. The superior performance of Model 1 (80/10/10 split) over Model 2 (70/15/15 split) underscores the importance of dataset allocation optimization, particularly with limited training data. However, both models exhibited class imbalance effects, achieving higher F1 scores for non-fracture detection (93.7-94%) compared to fracture detec-

 doi:10.20944/preprints202511.0873.v2

tion (69.2-81.6%). From a clinical perspective, this performance disparity is not recommended for deployment: a fracture F1 score of 69.2.9-81.6% implies that approximately 18-30% of fracture cases are either missed (false negatives) or incorrectly identified (false positives). Given that untreated or delayed treatment of elbow fractures can result in permanent complications including nerve palsy, joint stiffness, and cubitus varus [15], the clinical cost of false negatives far exceeds that of false positives. Current clinical standards would require fracture detection sensitivity exceeding 95% to ensure patient safety. The dataset's 2:1 non-fracture-to-fracture ratio likely contributed to this disparity. Future implementations must prioritize fracture detection sensitivity through weighted loss functions (assigning higher penalty to fracture misclassification), focal loss to address class imbalance, or stratified data augmentation specifically targeting the minority fracture class. Additionally, deployment safeguards such as confidence thresholding that favors sensitivity over specificity should be implemented to minimize false negatives.

Two models, model 1 and model 2, were tested for their ability to detect elbow fractures from radiographs. Model 1 displayed better performance over model 2, with a mean accuracy of 93.4%, F1 score of 92.7% and mAP@.5 of 70.4% when tested on scans of radiographs inside of the Android application.

Compared to the pre-export results, both models demonstrated higher F1 scores across both classes, while mAP@50 and mAP@50-95 values decreased. The increase in F1 appears to stem from a general improvement in precision and recall across both classes. This, in turn, most likely stems from differences in confidence score handling: YOLO models use a default confidence threshold of 25% for a detection to be considered successful, whereas in models 1 and 2 the threshold was 0%. This means that some low-confidence predictions, although correct, would have been misclassified by the YOLO model, whereas they would have passed in models 1 and 2. This is further evidenced by the mAP scores, which take into account both confidence and bounding box coordinates. These metrics have deteriorated in models 1 and 2 relative to the pre-export model - this behaviour could be explained by the introduction of NMS into the exported models. NMS affects which bounding boxes the model keeps, which can indirectly affect predicted bounding box coordinates and confidence scores, leading to a decrease in mAP.

These results for accuracy are comparable with the current state-of-the-art for elbow detection, with Li et al. achieving an accuracy of 94.1% [9]. Furthermore, the mAP@0.5 score can be compared to the current benchmark of 53.5% for wrist fracture detection, as achieved by Ferdi [28].

The model displayed acutely diminished performance when using the phone's camera for testing, with results heavily affected by camera positioning and lighting conditions. For fracture detection models to successfully detect fractures through this method, either by capturing photographs or using live detection, a high-quality, robust dataset created with this purpose is required.

The application's resource usage was relatively low on emulated Android devices, suggesting that such fracture detection applications are feasible hardware-wise. These findings establish that current mobile fracture detection systems face a fundamental tension: controlled digital radiograph inference achieves clinical-grade accuracy ($F1 = 92.7\%$), but practical deployment scenarios (photographed films, variable lighting, non-professional positioning) induce 4.4-9.8$\times$ error amplification that renders the system unreliable. Safe deployment requires either: (1) PACS integration to provide direct digital radiograph access, bypassing the camera acquisition pathway entirely, or (2) substantial investment in environmental compensation mechanisms that treat acquisition conditions as measurable, compensable covariates rather than noise to be ignored. The latter approach, while technically feasible following Liu et al.'s precedent, would require condition-aware training datasets 10-20$\times$ larger than the current 1,100-image corpus to adequately sample the environmental state space.

5.2. Generalisation Challenges

The confidence interval plots from Figure 4 reveal several critical insights about model performance across fracture detection tasks. **Model 1** consistently outperforms **Model 2** across both classes and all IoU thresholds, with particularly pronounced advantages in fracture detection (starting

 doi:10.20944/preprints202511.0873.v2

at $0.60mAP$ vs $0.47mAP$ at $IoU0.5$). Notably, both models struggle more with fracture detection compared to non-fracture detection, with the NoFrac class consistently achieving $10-20\%$ higher mAP scores. This finding suggests the models find it harder to identify subtle abnormalities. The overlapping curves between $FP16$ and $FP32$ quantisation schemes confirm negligible performance degradation, providing strong evidence that $FP16$ deployment offers computational efficiency without sacrificing accuracy. All models show steep performance decline as IoU thresholds increase from 0.5 to 0.8, with performance approaching zero by IoU 0.9, indicating that very strict localisation requirements are challenging for both models. A critical limitation of independent validation was the inability to evaluate spatial localisation metrics (mAP, IoU) due to annotation incompatibility. This restricts conclusions to classification accuracy rather than comprehensive detection performance, limiting assessment of the model's clinical utility for precise fracture localisation—a key advantage of object detection over binary classification.

The performance comparisons between in-app and independent testing revealed significant drops as indicated in Table 7 in subsection 4.2.1 and backed-up by a strong inverse relationship from Spearman rank test ($\rho = -0.866, p = 0.333$) with no statistical significance. This complete ranking instability suggests that performance on the in-app test set is a poor predictor of generalisation capability, with potential explanations including:

- Overfitting to training distribution: Model 1's aggressive 80/10/10 split may have optimised for in-distribution performance at the cost of generalisation.
- Dataset bias: The in-app test set may share systematic characteristics with training data (same institution, imaging protocols, patient demographics) that are absent in the independent Roboflow dataset.
- Model regularisation trade-offs: Model 2's lower in-app accuracy may reflect beneficial underfitting that translates to better cross-dataset robustness.

While statistical power limitations preclude definitive conclusions (p = 0.333), the magnitude of rank reversal (Model 2: +2 rank improvement, Model 1-FP32: -1.5 rank decline) represents a clinically significant finding that challenges conventional model selection based solely on held-out test set performance.

From a practical standpoint, Model 1 should be prioritized for deployment given its superior fracture detection capabilities, and FP16 quantization can be safely adopted for efficiency gains. The wider confidence intervals observed in Model 2, particularly at lower IoU thresholds, suggest less stable predictions compared to Model 1. The consistent underperformance in the fracture class across all configurations highlights a critical area for improvement, as lower fracture sensitivity could lead to missed diagnoses with serious clinical consequences. Future work should focus on addressing this class imbalance through targeted data augmentation, class-weighted loss functions, or ensemble methods specifically designed to boost fracture detection performance, while operating at IoU thresholds around $0.5-0.6$ where the models demonstrate optimal balance between precision and localization accuracy.

Independent testing revealed concerning variability in model performance. Model 2 achieved an insignificant gain of 66.3% accuracy on the indepndent Roboflow test set compared to Model 1's 64%; a reversal of their relative performance on the original test set. While the independent set was relatively small (n=214), this finding suggests potential overfitting to specific characteristics of the training dataset, such as image quality, contrast levels, or institutional imaging protocols. This aligns with research by Kutbi [2] and Zech et al. [30] emphasizing that diverse, multi-institutional datasets are essential for robust clinical AI systems. The models' vulnerability to distribution shifts underscores the need for training data that captures real-world heterogeneity across imaging equipment, exposure settings, patient positioning, and image processing pipelines.

The findings from section 4.5.1 requires future architectural and training modifications focused around:

1. Replacing standard BCE with focal loss [9] using $\gamma =$ 2-3 to down-weight well-classified examples and focus optimization on hard fracture cases

 doi:10.20944/preprints202511.0873.v2

2. Implement cost-sensitive weighting mechanism with $w_1 = c_{fn}/c_{fp} \approx$ 10-20 to explicitly penalize false negatives during training
3. Threshold Optimization through post-hoc calibration to reduce classification threshold from default 0.5 to ≈0.3, deliberately increasing false positive rate to achieve fracture recall $> 95\%$
4. Utilizing Ensemble Uncertainty by deploying multiple model variants with uncertainty quantification to flag low-confidence predictions for mandatory radiologist review

Our results empirically validate the theoretical insight that domain-agnostic robustness cannot be achieved through training data diversity alone. Even with horizontal flip augmentation (effectively $2\times$ training set), contrast stretching, and extensive hyperparameter optimization (60 Optuna trials), the models failed to learn invariances spanning the digital-to-camera domain gap. The magnitude of $d_{\mathcal{H}\Delta\mathcal{H}} \approx 0.43$ implies that no hypothesis in the current model class (YOLOv11 with ImageNet pretraining) can simultaneously minimize both $R_{train}(f)$ and $R_{camera}(f)$ without explicit environmental modeling.

5.3. Performance Degradation as Domain Shift: Environmental Compensation Requirements

Interesting observations encompass the performance degradation of the model in multiple aspects that are worth exploring further. For example, the decline of performance between mAP50 and mAP50-95 outlined in Table 3. This can be attributed to the increased technical "stringency" of the $mAP50-95$ metric, which averages precision across ten incremental Intersection over Union (IoU) thresholds. While the model achieves high diagnostic accuracy at the more lenient 0.5 threshold, indicating it can successfully identify the presence of a fracture, the sharp drop in the averaged metric highlights a specific deficiency in the model's spatial localization capability, as it struggles to maintain the sharpness of the bounding box alignment required for higher IoU levels like 0.95. This performance gap is scientifically attributed to the inherent difficulty of precisely delineating subtle bone fractures, a task further complicated by dataset limitations, such as class imbalance and a lack of diverse viewing conditions. Additionally, the sensitivity of the YOLOv11 architecture to potential "annotation noise" in the training labels suggests that even minor inconsistencies in manual labeling can significantly penalize stricter metrics, demonstrating that while the model is effective for general detection, it currently lacks the geometric precision necessary for high-fidelity clinical localization.

The pronounced degradation in model performance observed when transitioning from digital radiographs to camera-based inference (with F1 scores decreasing from 92.7% to 31–60.3%) can be interpreted as a domain shift problem that necessitates systematic environmental compensation strategies. This phenomenon is consistent with documented challenges in structural health monitoring, where temperature-induced variations in ultrasonic guided wave systems result in temporal shifts and amplitude attenuation, thereby generating false positives even in the absence of genuine structural damage.[38]. Environmental factors that systematically distort sensor inputs without touching the actual phenomenon of interest demand careful compensation, not assumptions about invariant data distributions. Liu et al. (2025) showed that simple temperature shifts can change material properties and signal behaviour enough to push monitoring signals far from their baselines, even when the structure itself is perfectly healthy.

Our fracture detection system faces the same problem: shifts in lighting, camera angle, surface reflections, or lens distortion can radically change how a radiograph appears, while the underlying presence or absence of a fracture remains unchanged. The input distribution changes; the ground truth does not, and this can be tackled by explicitly modelling the environment. Liu et al. (2025) built multi-temperature baseline libraries and used a two-stage compensation pipeline: a coarse step to select the optimal baseline, followed by a fine Hilbert-domain correction. By engineering around environmental variation instead of simply trusting model robustness, they hit 99.43% accuracy across a 70°C temperature swing—a clear demonstration that explicit compensation beats blind generalization.

Our findings indicate that robust clinical deployment of camera-based fracture detection requires analogous environmental adaptation strategies. Just as Liu et al. could not achieve reliable struc-

 doi:10.20944/preprints202511.0873.v2

tural monitoring without temperature compensation—even with sophisticated signal processing; our models cannot reliably interpret photographed radiographs without systematic compensation for acquisition conditions. Potential approaches include:

1. Constructing "environmental baseline libraries" containing radiographs captured under systematically varied lighting conditions, angles, and distances;
2. Implementing preprocessing pipelines that normalize lighting and perspective variations before inference;
3. Developing adaptive inference mechanisms that detect environmental conditions (via metadata or auxiliary networks) and apply condition-specific correction factors;
4. Training models with explicit environmental augmentation that treats lighting and perspective as measurable covariates rather than noise to be ignored. In our case, the 28.8-43.1% F1 scores for live detection, despite 92.7% performance on curated datasets, exemplify this principle.

Future work must move beyond hoping for generalization to engineering systematic compensation mechanisms that account for the "operating envelope" of environmental conditions under which the system will function clinically.

5.4. Computational Efficiency

A major contribution of this research is demonstrating that YOLOv11 models operate efficiently on mobile devices with modest hardware. Resource usage remained conservative: static image inference consumed under $0.2GB$ RAM and less than 4% CPU, while live detection required approximately $0.25G - 0,29$ RAM and 24% CPU. These demands suggest the application could function on budget Android devices with 4GB RAM, addressing accessibility concerns in resource-limited settings. FP16 quantization proved particularly valuable, reducing model size by 50% while producing virtually identical performance to $FP32$ versions across all tests. The minimal difference (e.g., Model 1's $mAP@50$ of 70.4% for FP16 vs. 69.3% for FP32) validates aggressive model compression for mobile deployment. This computational efficiency, combined with YOLO's single-stage detection architecture providing faster inference than two-stage detectors like Faster R-CNN, positions YOLOv11 as a viable architecture for real-time mobile medical imaging applications—if the domain shift challenges can be resolved.

5.5. Ethical Considerations

Beyond technical performance, mobile AI fracture detection raises significant ethical concerns requiring careful consideration before clinical deployment. The system's current fracture detection f1-scores ($69.2 - 81.6\%$) and performance degradation under real-world camera conditions (F1: 28.8-60.3%) highlight critical patient safety implications: deploying inadequately validated systems risks missed diagnoses with serious consequences including permanent joint damage and nerve complications [15]. Accountability frameworks remain undefined when errors occur, responsibility distribution among AI developers, healthcare providers, and institutions lacks clear legal precedent. Health equity concerns are particularly acute: while mobile AI promises improved access in underserved regions, our findings that performance depends on environmental conditions and device quality may paradoxically worsen disparities if resource-limited settings lack optimal imaging conditions or high-quality smartphones. Additionally, informed consent and transparency requirements mandate that patients understand AI's role in their care, including specific error rates and limitations, yet communicating these technical complexities in an accessible manner remains challenging. Finally, data privacy and security considerations demand that any mobile health application handling radiographic images implement robust encryption and comply with regulations, including GDPR and HIPAA, with explicit patient consent for data collection, storage, and potential model improvement. These ethical dimensions must be systematically addressed through multi-stakeholder engagement involving clinicians, ethicists, patients, regulators, and AI developers before widespread implementation.

 doi:10.20944/preprints202511.0873.v2

Furthermore, the clinical requirement for the deployed model would require regulatory approval as a Class II medical device as per US FDA and EU MDR regulations. Key regulatory requirements include: (1) prospective multi-center validation studies with radiologist ground truth, (2) demonstration of fracture sensitivity >95% per FDA guidance on computer-aided detection, (3) Failure Mode and Effects Analysis (FMEA) addressing false negative scenarios, and (4) post-market surveillance for performance monitoring. Current fracture recall rates (59.1–92.3%, Table 11) fall below regulatory thresholds, necessitating class-weighted training and threshold optimisation before regulatory submission.

6. Conclusions

The research makes important contributions: (1) demonstrating YOLOv11 can achieve clinically relevant accuracy while maintaining mobile-friendly computational requirements; (2) identifying the critical gap between laboratory validation and real-world camera-based performance; (3) quantifying environmental factors' impact on model reliability; and (4) establishing that current training paradigms are insufficient for robust camera-based inference. Key strengths include computational efficiency, successful quantization, competitive accuracy on curated datasets, and systematic evaluation exposing critical limitations. Principal limitations center on inadequate dataset diversity, severe camera-based performance degradation, clinically inadequate fracture detection sensitivity due to class imbalance (with false negative rates undesirable for patient safety), and absence of clinical validation with real patient outcomes For reliable smartphone-based fracture detection, future work must develop diverse datasets explicitly incorporating varied viewing conditions, lighting scenarios, and acquisition contexts. Alternative approaches such as direct PACS integration or hybrid systems with uncertainty-aware inference may prove more viable. While this research demonstrates that AI-powered mobile fracture detection is technically feasible under controlled conditions, substantial innovations in domain-aware training, robust uncertainty quantification, and prospective clinical validation are required before such systems can safely assist clinicians in real-world settings. The findings emphasize that impressive laboratory metrics do not guarantee clinical utility when deployment conditions fundamentally differ from training environments, providing crucial insights for future medical AI development. It is worth noting that this study evaluated technical performance without clinician interaction. Critical unanswered questions, such as trust calibration when predictions conflict with clinical judgment, and whether the system truly enhances diagnostic efficiency in time-pressured emergency settings, remain. Prospective clinical trials involving practicing physicians are essential before deployment.

Future research should prioritize the following crucial aspects:

1. Developing diverse datasets explicitly incorporating camera-acquired images under varied conditions;
2. Implementing domain adaptation techniques like CycleGAN to bridge digital-to-photograph gaps;
3. Creating specific models addressing anatomical differences;
4. Implementing the cost-sensitive framework from equation 10 with empirical clinical cost ratios derived from elbow fracture outcome studies, and conducting ablation studies comparing BCE, weighted BCE, and focal loss under identical conditions.
5. Exploring hybrid approaches combining PACS integration for clinical settings with camera functionality for field use; and
6. Conducting prospective clinical trials comparing AI-assisted diagnosis with standard practice.

This study demonstrates that YOLOv11-based elbow fracture detection is technically feasible on mobile devices under controlled conditions, but substantial gaps separate laboratory performance from safe clinical deployment. Current results establish feasibility benchmarks and identify critical barriers rather than validating deployment-ready systems. The $4.4 - 9.8\times$ error amplification under real-world acquisition conditions, inadequate fracture sensitivity ($69.2 - 81.6\%$ F1), and absence of

 doi:10.20944/preprints202511.0873.v2

clinical workflow validation preclude near-term implementation without addressing the fundamental limitations identified in Sections 4.5 and 5.

Author Contributions: F.S has contributed to the data curation related to data analysis, results interpretation, and the investigation process, specifically performing the experiments and data/evidence collection. B.R has contributed to verifying the overall reproducibility of results and other research outputs, and to preparing and creating the published work, specifically critical reviews, commentaries, or revisions, including pre- and post-publication stages.

Funding: This research received no external funding.

Data Availability Statement: The datasets presented in this article are not readily available because it has been made private by its original authors. At this time there is no URL or any available party to make a request for this data.

Conflicts of Interest: The authors declare no conflicts of interest.

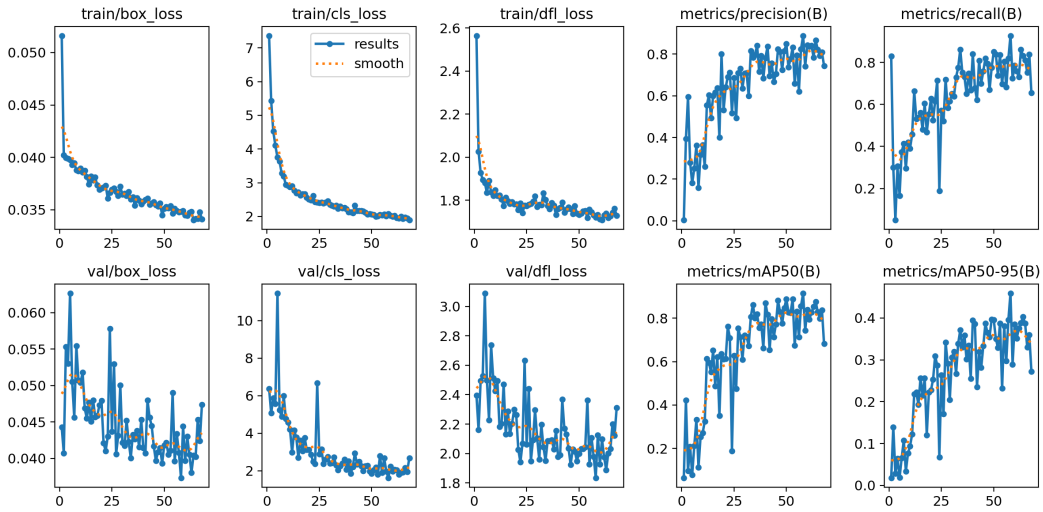
Abbreviations

The following abbreviations are used in this manuscript:

YOLO	You Only Look Once
CNN	Convolutional Neural Network
NMS	Non-Maximum Suppression
IoU	Intersection over Union
TFLite	TensorFlow Lite
FP	Floating Point
PR	Precision-Recall
mAP	mean Average Precision
PACS	Picture Archiving and Communication System

Appendix A. Training Curves for Model 1 and Model 2

Appendix A.1. Training and Validation Curves for Model 1

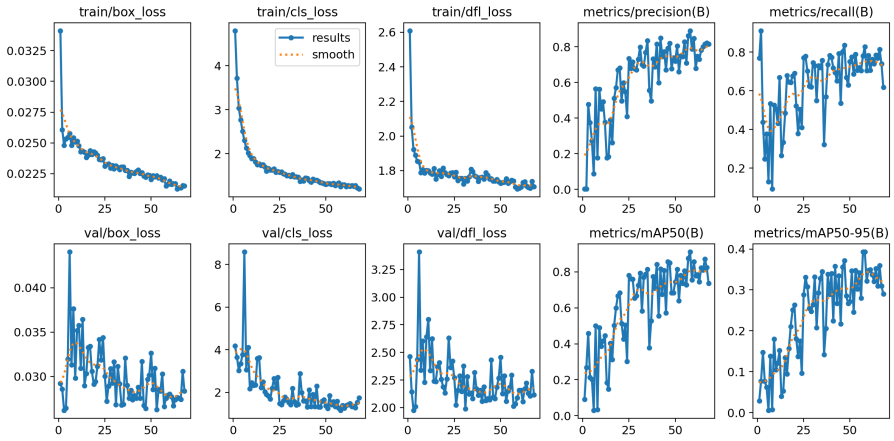


(**a**)

Figure A1. The figure illustrates subplots showing the progression of various loss functions and performance metrics during model training. The top row shows training losses: bounding box regression loss (box_loss), localization loss (ls_loss), and distribution focal loss (dfl_loss), all demonstrating steady convergence from initial high values to stable low values around epoch 50-70. The middle row presents validation losses for the same metrics, with box_loss and cls_loss showing good convergence, though with more variability than training losses,

 doi:10.20944/preprints202511.0873.v2

while dfl_loss exhibits substantial noise throughout validation. The bottom row displays object detection performance metrics: precision and recall (both around 0.8 by epoch 70), mean Average Precision at IoU threshold 0.5 (mAP50, reaching 0.85), and mAP50-95 (around 0.35). The orange dotted line in the $train/ls_loss$ plot represents smoothed values for better trend visualization. Overall, the metrics indicate successful model convergence with strong detection performance, though the noisy validation losses suggest some instability in the validation dataset or evaluation process.



(**a**)

Figure A2. The figure illustrates subplots showing the progression of various loss functions and performance metrics during model training. The top row shows training losses: bounding box regression loss (box_loss), localization loss (ls_loss), and distribution focal loss (dfl_loss), all demonstrating steady convergence from initial high values to stable low values around epoch 50-70. The middle row presents validation losses for the same metrics, with box_loss and cls_loss showing good convergence, though with more variability than training losses, while dfl_loss exhibits substantial noise throughout validation. The bottom row displays object detection performance metrics: precision and recall (both around 0.8 by epoch 70), mean Average Precision at IoU threshold 0.5 (mAP50, reaching 0.75), and mAP50-95 (around 0.35). The orange dotted line in the $train/ls_loss$ plot represents smoothed values for better trend visualization. Overall, the metrics indicate successful model convergence with strong detection performance, though the noisy validation losses suggest some instability in the validation dataset or evaluation process.

Appendix B. The Results of Models for Camera and Live Detection

Table A1. Results for model inference using device's camera in natural light (FP32).

Camera Detection Results - Daylight (Model 1 FP32)				
Performance Metric	Straight Down	45 paper shift	45 side angle	average
Fracture detection precision	0.6	0.5	0.5	0.533
Fracture detection recall	0.6	0.6	0.8	0.667
Fracture detection F1 score	0.6	0.545	0.615	0.587
Non-Fracture detection precision	0.75	0.667	1.0	0.806
Non-Fracture detection recall	0.6	0.4	0.2	0.4
Non-Fracture detection F1 score	0.667	0.533	0.333	0.51
Average Precision	0.675	0.583	0.75	0.669
Average Recall	0.6	0.5	0.5	0.533
Average F1 Score	0.634	0.539	0.474	0.549

 doi:10.20944/preprints202511.0873.v2

Table A2. Results for model inference using device's camera in artificial light (FP32).

Camera Detection Results - Artificial Light (Model 1 FP32)				
Performance Metric	Straight Down	45 paper shift	45 side angle	average
Fracture detection precision	0.6	0.5	0.429	0.51
Fracture detection recall	0.6	0.4	0.6	0.533
Fracture detection F1 score	0.6	0.444	0.5	0.515
Non-Fracture detection precision	0.6	0.75	0.5	0.617
Non-Fracture detection recall	0.6	0.6	0.2	0.467
Non-Fracture detection F1 score	0.6	0.667	0.286	0.518
Average Precision	0.6	0.625	0.465	0.563
Average Recall	0.6	0.5	0.4	0. 5
Average F1 Score	0.6	0.555	0. 393	0.516

Table A3. Results for model inference using device's camera in natural light (FP16).

Camera Detection Results - Daylight (Model 1 FP16)				
Performance Metric	Straight Down	45 paper shift	45 side angle	average
Fracture detection precision	0.75	0.75	0.5	0.667
Fracture detection recall	0.6	0.6	0.2	0.467
Fracture detection F1 score	0.667	0.667	0.286	0.54
Non-Fracture detection precision	0.667	0.75	0.5	0.639
Non-Fracture detection recall	0.8	0.75	0.8	0.783
Non-Fracture detection F1 score	0.632	0.75	0.615	0.666
Average Precision	0.708	0.75	0.5	0.653
Average Recall	0.7	0.675	0.5	0.625
Average F1 Score	0.649	0.708	0. 451	0.603

Table A4. Results for model inference using device's camera in artificial light (FP16).

Camera Detection Results - Artificial Light (Model 1 FP16)				
Performance Metric	Straight Down	45 paper shift	45 side angle	average
Fracture detection precision	0.5	0.25	0.333	0.361
Fracture detection recall	0.6	0.2	0.4	0.4
Fracture detection F1 score	0.545	0.222	0.364	0.377
Non-Fracture detection precision	0.333	0.333	0.25	0.306
Non-Fracture detection recall	0.2	0.2	0.2	0.2
Non-Fracture detection F1 score	0.25	0.25	0.222	0.241
Average Precision	0.417	0.292	0.292	0.334
Average Recall	0.4	0.2	0.3	0.3
Average F1 Score	0.398	0.236	0.293	0.31

 doi:10.20944/preprints202511.0873.v2

Table A5. Results for live model inference in natural light (FP32).

Live Camera Detection Results - Daylight (Model 1 FP32)				
Performance Metric	Straight Down	45 paper shift	45 side angle	average
Fracture detection precision	0.333	0.625	0.333	0.431
Fracture detection recall	0.4	1.0	0.4	0.6
Fracture detection F1 score	0.364	0.769	0.364	0.499
Non-Fracture detection precision	0.25	1.0	0.333	0.528
Non-Fracture detection recall	0.2	0.4	0.2	0.267
Non-Fracture detection F1 score	0.222	0.571	0.25	0.348
Average Precision	0.292	0.813	0.333	0.479
Average Recall	0.3	0.7	0.3	0.433
Average F1 Score	0.293	0.67	0.307	0.423

Table A6. Results for live model inference in natural light (FP16).

Live Camera Detection Results - Daylight (Model 1 FP16)				
Performance Metric	Straight Down	45 paper shift	45 side angle	average
Fracture detection precision	0.25	0.625	0.333	0.403
Fracture detection recall	0.2	1.0	0.4	0.533
Fracture detection F1 score	0.222	0.769	0.364	0.452
Non-Fracture detection precision	0.2	1.0	0.333	0.511
Non-Fracture detection recall	0.2	0.4	0.2	0.267
Non-Fracture detection F1 score	0.2	0.571	0.25	0.340
Average Precision	0.225	0.813	0.333	0.457
Average Recall	0.2	0.7	0.3	0.4
Average F1 Score	0.211	0.67	0.307	0.396

Table A7. Results for live model inference in artificial light (FP32).

Live Camera Detection Results - Artificial Light (Model 1 FP32)				
Performance Metric	Straight Down	45 paper shift	45 side angle	average
Fracture detection precision	0.2	0.571	0.333	0.368
Fracture detection recall	0.2	0.8	0.4	0.467
Fracture detection F1 score	0.2	0.667	0.364	0.41
Non-Fracture detection precision	0.0	0.667	0.0	0.222
Non-Fracture detection recall	0.0	0.4	0.0	0.133
Non-Fracture detection F1 score	0.0	0.5	0.0	0.167
Average Precision	0.1	0.619	0.167	0.295
Average Recall	0.1	0.6	0.2	0.3
Average F1 Score	0.1	0.583	0.182	0.288

 doi:10.20944/preprints202511.0873.v2

Table A8. Results for live model inference in artificial light (FP16).

Live Camera Detection Results - Artificial Light (Model 1 FP16)				
Performance Metric	Straight Down	45 paper shift	45 side angle	average
Fracture detection precision	0.333	0.667	0.5	0.5
Fracture detection recall	0.5	0.8	0.4	0.567
Fracture detection F1 score	0.4	0.727	0.444	0.524
Non-Fracture detection precision	0.0	1.0	0.5	0.5
Non-Fracture detection recall	0.0	0.4	0.4	0.267
Non-Fracture detection F1 score	0.0	0.571	0.444	0.338
Average Precision	0.167	0.833	0.5	0.5
Average Recall	0.25	0.6	0.4	0.417
Average F1 Score	0.2	0.649	0.444	0.431

Table A9. Post-training performance metrics of Models 1 & 2.

Model Performance Results		
Performance Metric	Model 1	Model 2
Fracture detection precision	0.701	0.826
Fracture detection recall	0.923	0.591
Fracture detection F1 score	0.797	0.689
Fracture detection mAP50	0.816	0.692
Fracture detection mAP50-95	0.348	0.266
Non-Fracture detection precision	0.942	0.874
Non-Fracture detection recall	0.857	0.852
Non-Fracture detection F1 score	0.897	0.862
Non-Fracture detection mAP50	0.94	0.937
Non-Fracture detection mAP50-95	0.467	0.483
Average precision	0.821	0.85
Average recall	0.89	0.722
Average mAP50	0.878	0.814
Average mAP50-95	0.408	0.375
Average F1 Score	0.854	0.781

 doi:10.20944/preprints202511.0873.v2

Table A10. In-app model performance.

Model Performance Results				
Performance Metric	model 1 - FP32	model 1 - FP16	model 2 - FP32	model 2 - FP16
Fracture detection precision	0.846	0.846	0.893	0.893
Fracture detection recall	0.971	0.971	0.877	0.877
Fracture detection F1 score	0.904	0.904	0.885	0.885
Fracture average mAP@50	0.65	0.65	0.51	0.51
Fracture average mAP@50-95	0.310	0.312	0.223	0.223
Non-Fracture detection precision	0.985	0.985	0.936	0.936
Non-Fracture detection recall	0.917	0.917	0.945	0.945
Non-Fracture detection F1 score	0.95	0.95	0.941	0.941
Non-Fracture average mAP@50	0.736	0.757	0.79	0.79
Non-Fracture average mAP@50-95	0.373	0.374	0.386	0.386
Average F1 Score	0.927	0.927	0.913	0.913
Average Accuracy	0.934	0.934	0.922	0.922
Average Bounding Box IoU	0.699	0.699	0.683	0.683
Average mAP@50	0.693	0.704	0.65	0.65
Average mAP@50-95	0.342	0.343	0.305	0.305
Average Confidence	0.698	0.698	0.676	0.676

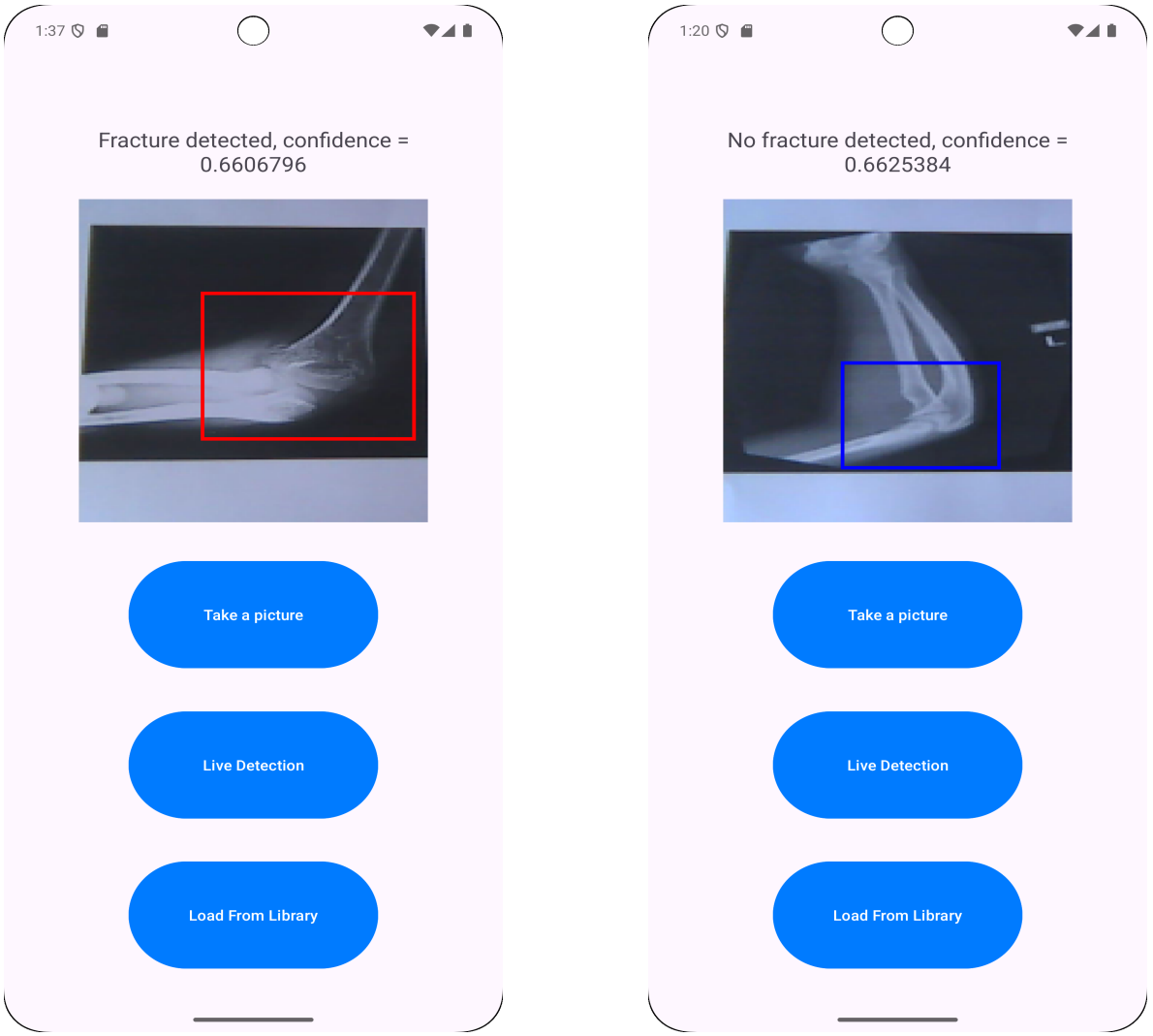
Table A11. Independent testing results for Models 1 & 2.

Model Performance Results - Independent Test Set		
Performance Metric	Model 1 FP32 & FP16	Model 2 FP32 & FP16
Fracture detection precision	0.667	1.0
Fracture detection recall	0.8	0.8
Fracture detection F1 score	0.727	0.889
Non-Fracture detection precision	0.75	0.833
Non-Fracture detection recall	0.6	1.0
Non-Fracture detection F1 score	0.667	0.91
Average F1 Score	0.697	0.889
Average Confidence	0.682	0.625
Average Accuracy	0.7	0.9

 doi:10.20944/preprints202511.0873.v2

Appendix C. Images for Correctly and Incorrectly Classified Cases from Image Radiographs and Live Predictions

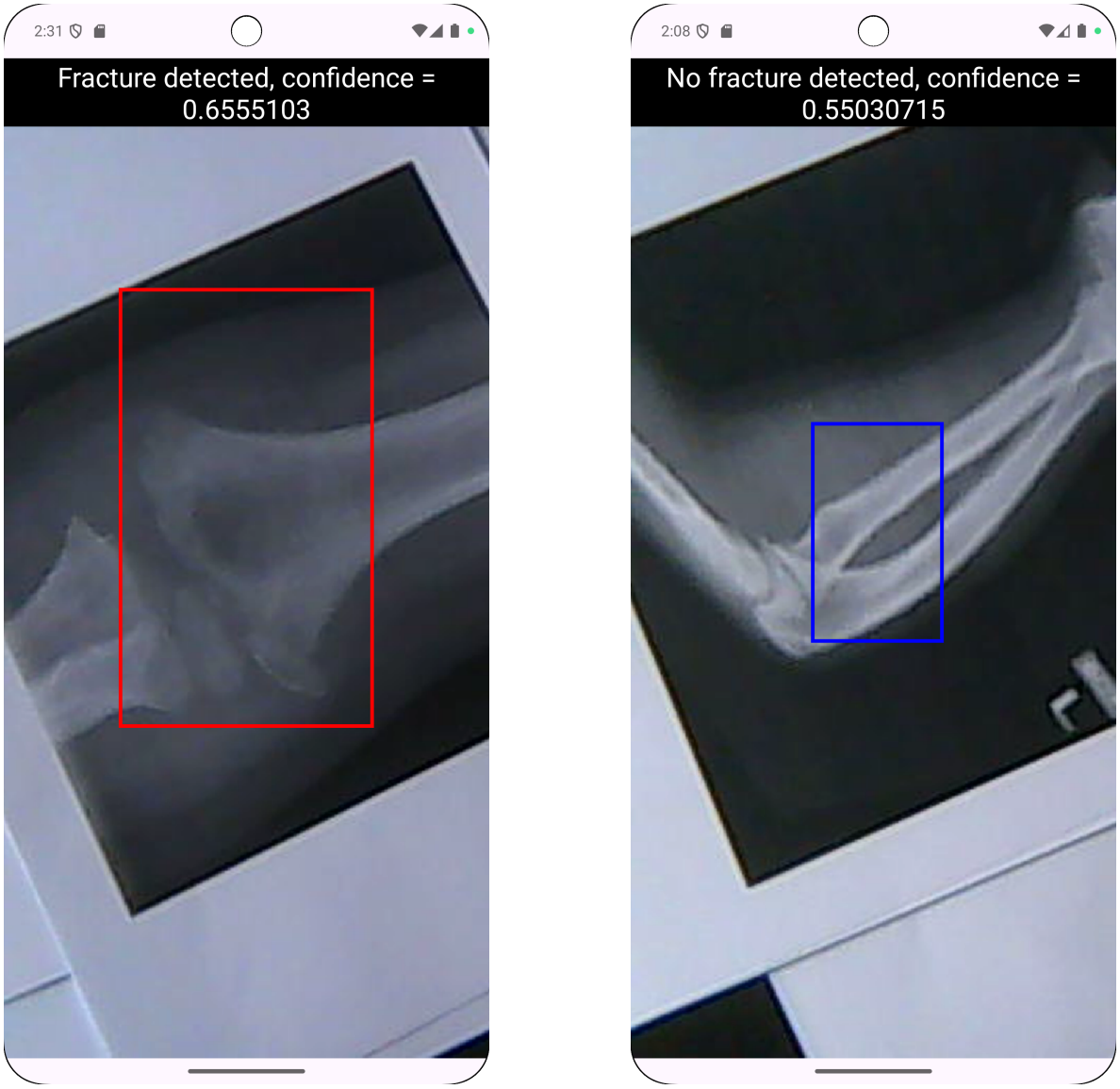
Appendix C.1. Images of Correctly Classified Cases



(**a**) (**b**)

Figure A3. Example of correct classifications of radiograph photographs: (**a**) Fracture detected on radiograph photo. (**b**) Non-fracture detected on radiograph photo.

 doi:10.20944/preprints202511.0873.v2

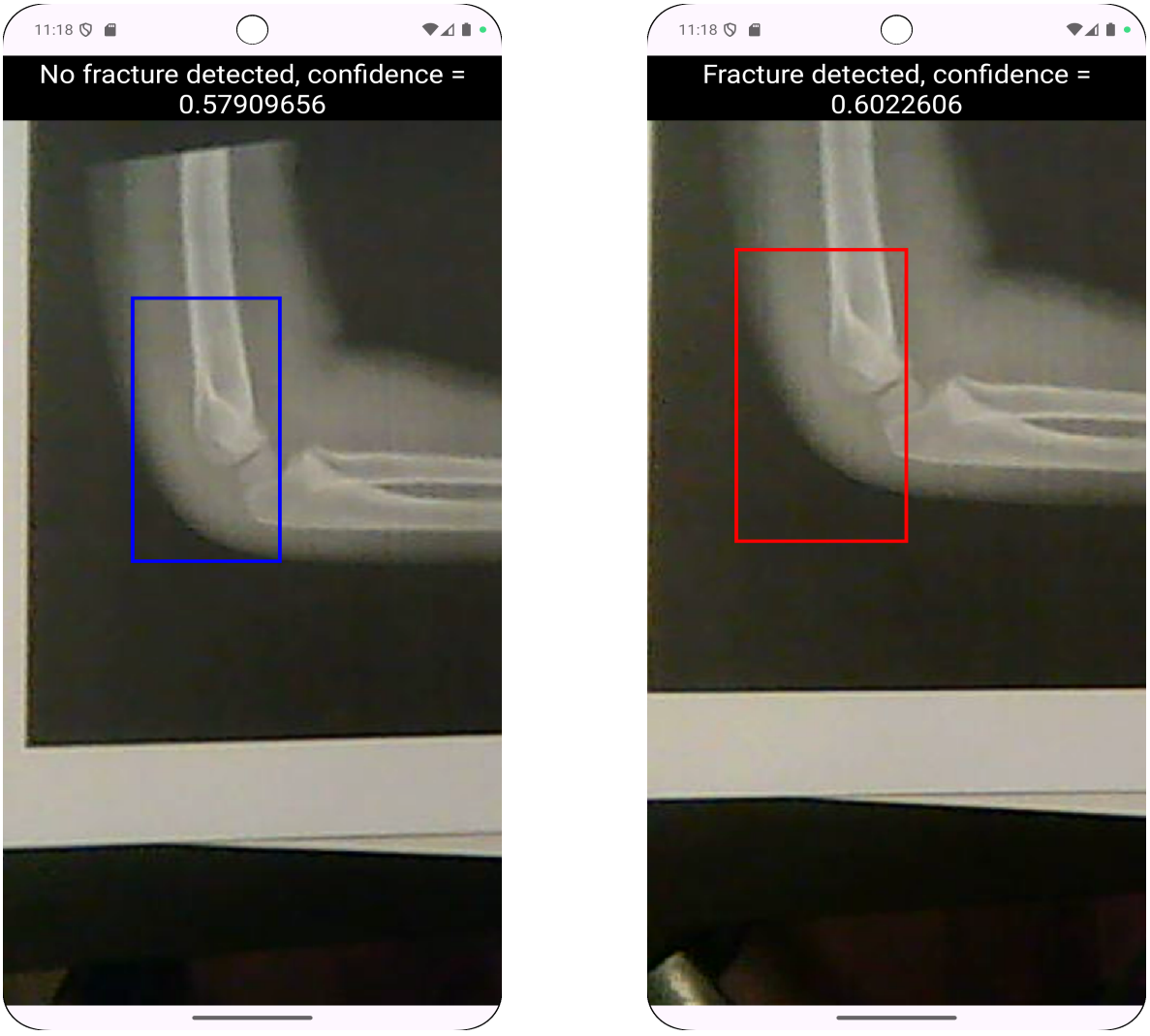


(**a**) (**b**)

Figure A4. Examples of successful live predictions: (**a**) Fracture detected with live detection. (**b**) Non-fracture detected with live detection.

 doi:10.20944/preprints202511.0873.v2

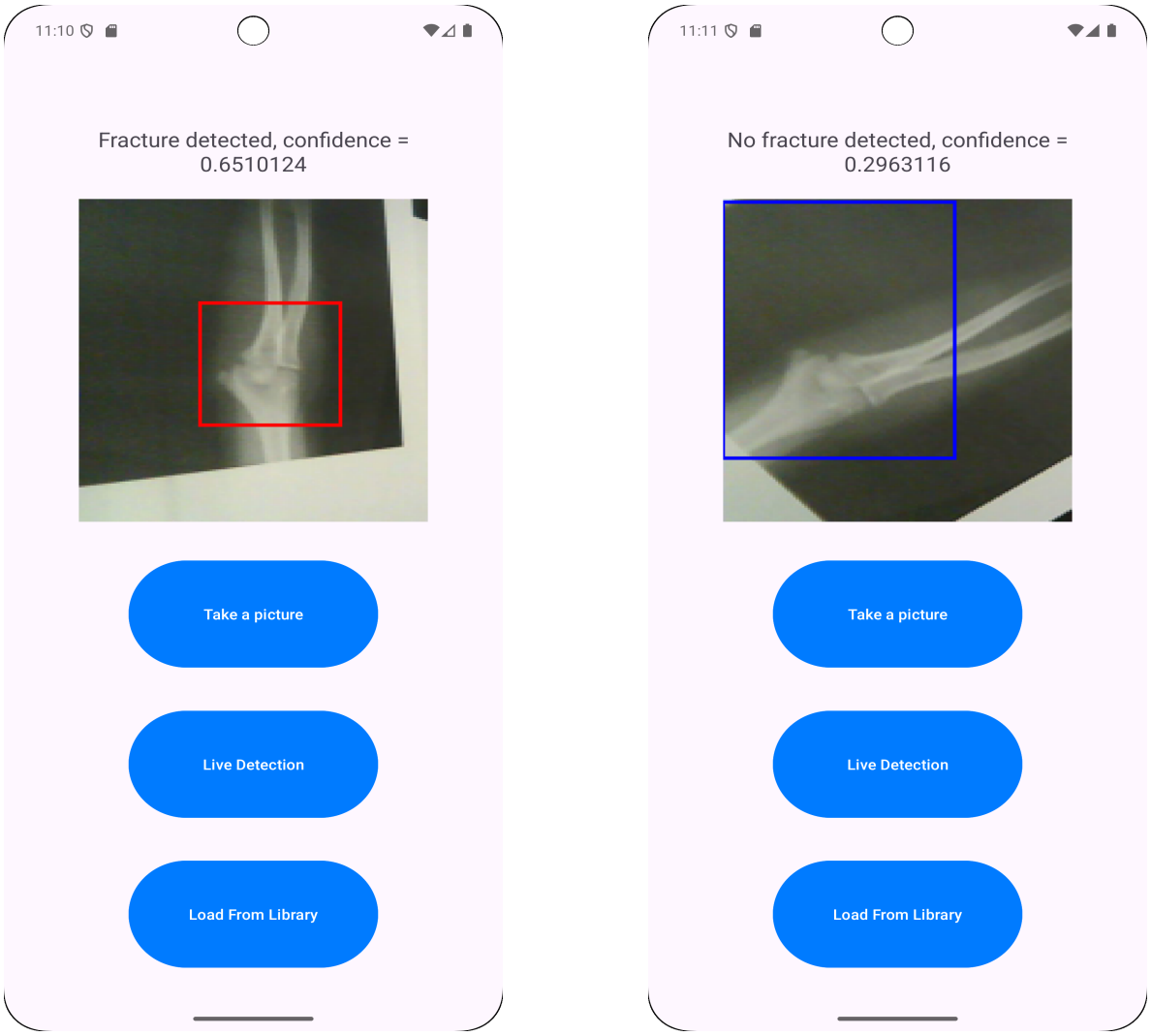
Appendix C.2. Images of Incorrectly Classified Cases



(**a**) (**b**)

Figure A5. Example showing live detection inconsistency in the same lighting conditions and viewed from the same angle. This was counted as a false negative for the class in the image.

 doi:10.20944/preprints202511.0873.v2



(**a**) (**b**)

Figure A6. Example of prediciton inconsistency at various angles: (**a**) Fracture detected on radiograph photo. (**b**) Non-fracture detected on the same radiograph photo.

References

1. Kuo, R.Y.L.; Harrison, C.; Curran, T.; Jones, B.; Freethy, A.; Cussons, D.; Stewart, M.; Collins, G.; Furniss, D. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. *Radiology* **2022**, p. 211785. https://doi.org/10.1148/radiol.211785.
2. Kutbi, M. Artificial Intelligence-Based Applications for Bone Fracture Detection Using Medical Images: A Systematic Review. *Diagnostics* **2024**, *14*. https://doi.org/10.3390/diagnostics14171879.
3. Nurifin, S. Performance of artificial intelligence in detecting bone fractures in radiographic results: A systematic literature review. *Malahayati International Journal of Nursing and Health Science* **2025**. https://doi.org/10.33024/minh.v8i1.666.
4. Mishra, A.K. Evolution of Artificial Intelligence in Bone Fracture Detection. *International Journal of Reliable and Quality E-Healthcare* **2022**. https://doi.org/10.4018/ijrqeh.299958.
5. Guermazi, A.; Tannoury, C.; Kompel, A.; Murakami, A.M.; Ducarouge, A.; Gillibert, A.; Li, X.; Tournier, A.; Lahoud, Y.; Jarraya, M.; et al. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. *Radiology* **2021**, p. 210937. https://doi.org/10.1148/radiol.210937.
6. Bousson, V.; Attané, G.; Benoist, N.; Perronne, L.; Diallo, A.; Hadid-Beurrier, L.; Martin, E.; Hamzi, L.; Duval, A.D.; Revue, E.; et al. Artificial Intelligence for Detecting Acute Fractures in Patients Admitted to an Emergency Department: Real-Life Performance of Three Commercial Algorithms. *Academic radiology* **2023**. https://doi.org/10.1016/j.acra.2023.06.016.

 doi:10.20944/preprints202511.0873.v2

7. Canoni-Meynet, L.; Verdot, P.; Danner, A.; Calame, P.; Aubry, S. Added value of an artificial intelligence solution for fracture detection in the radiologist's daily trauma emergencies workflow. *Diagnostic and interventional imaging* **2022**. https://doi.org/10.1016/j.diii.2022.06.004.
8. Dann, L.; Edwards, S.; Hall, D.; Davis, T.; Roland, D.; Barrett, M. Black and white: how good are clinicians at diagnosing elbow injuries from paediatric elbow radiographs alone? *Emerg Med J* **2024**, *41*, 662–667.
9. Li, J.; Hu, W.; Wu, H.; Chen, Z.; Chen, J.; Lai, Q.; Wang, Y.; Li, Y. Detection of hidden pediatric elbow fractures in X-ray images based on deep learning. *Journal of Radiation Research and Applied Sciences* **2024**, *17*, 100893. https://doi.org/https://doi.org/10.1016/j.jrras.2024.100893.
10. Boissin, C.; Blom, L.; Wallis, L.; Laflamme, L. Image-based teleconsultation using smartphones or tablets: qualitative assessment of medical experts. *Emergency Medicine Journal* **2017**, *34*, 95–99, [https://emj.bmj.com/content/34/2/95.full.pdf]. https://doi.org/10.1136/emermed-2015-205258.
11. M, P.P.; M, S.H.; N, R.; S, S.B. Edge AI-based Bone Fracture Detection using TFlite. *International Journal of Innovative Research in Advanced Engineering* **2025**. https://doi.org/10.26562/ijirae.2025.v1204.04.
12. Rajpurkar, P.; Lungren, M.P. The Current and Future State of AI Interpretation of Medical Images. *New England Journal of Medicine* **2023**, *388*, 1981–1990, [https://www.nejm.org/doi/pdf/10.1056/NEJMra2301725]. https://doi.org/10.1056/NEJMra2301725.
13. Pinto-Coelho, L. How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications. *Bioengineering* **2023**, *10*. https://doi.org/10.3390/bioengineering10121435.
14. Rayan, J.; Reddy, N.; Kan, J.; Zhang, W.; Annapragada, A. Binomial Classification of Pediatric Elbow Fractures Using a Deep Learning Multiview Approach Emulating Radiologist Decision Making. *Radiology: Artificial Intelligence* **2019**, *1*, e180015. https://doi.org/10.1148/ryai.2019180015.
15. Binh, L.N.; Nhu, N.T.; Nhi, P.T.U.; Son, D.L.H.; Bach, N.; Huy, H.Q.; Le, N.Q.K.; Kang, J.H. Impact of deep learning on pediatric elbow fracture detection: a systematic review and meta-analysis. *European Journal of Trauma and Emergency Surgery* **2025**, *51*, 115.
16. Oyeniyi, J.; Oluwaseyi, P. Emerging Trends in AI-Powered Medical Imaging: Enhancing Diagnostic Accuracy and Treatment Decisions. *International Journal of Enhanced Research In Science Technology & Engineering* **2024**, *13*, 2319–7463. https://doi.org/10.55948/IJERSTE.2024.0412.
17. Bhatnagar, A.; Kekatpure, A.L.; Velagala, V.R.; Kekatpure, A. A Review on the Use of Artificial Intelligence in Fracture Detection. *Cureus* **2024**, *16*, e58364.
18. Jacques, T.; Cardot, N.; Ventre, J.; Demondion, X.; Cotten, A. Commercially-available AI algorithm improves radiologists' sensitivity for wrist and hand fracture detection on X-ray, compared to a CT-based ground truth. *European Radiology* **2024**, *34*, 2885–2894.
19. Bhatnagar, A.; Kekatpure, A.L.; Velagala, V.R.; Kekatpure, A. A review on the use of artificial intelligence in fracture detection. *Cureus* **2024**, *16*.
20. Dupuis, M.; Delbos, L.; Rouquette, A.; Adamsbaum, C.; Veil, R. External validation of an artificial intelligence solution for the detection of elbow fractures and joint effusions in children. *Diagnostic and Interventional Imaging* **2024**, *105*, 104–109. https://doi.org/https://doi.org/10.1016/j.diii.2023.09.008.
21. Afacan, M.A.; Kılıç, K.; Temiz, A.; Tayfur, I.; Doganay, F. Diagnostic accuracy of fat pad sign, X-ray, and computed tomography in elbow trauma: implications for treatment choices—a retrospective study. *PeerJ* **2025**, *13*. https://doi.org/10.7717/peerj.18922.
22. Kargl, S.; Pumberger, W.; Luczyński, S.; Moritz, T. Assessment of interpretation of paediatric skeletal radiographs in the emergency room. *Clinical radiology* **2019**, *74 2*, 150–153. https://doi.org/10.1016/j.crad.2018.06.024.
23. Loeffen, D.V.; Zijta, F.; Boymans, T.A.; Wildberger, J.E.; Nijssen, E. AI for fracture diagnosis in clinical practice: Four approaches to systematic AI-implementation and their impact on AI-effectiveness. *European journal of radiology* **2025**, *187*, 112113. https://doi.org/10.1016/j.ejrad.2025.112113.
24. Takapautolo, J.; Neep, M.; Starkey, D. Analysing false-positive errors when Australian radiographers use preliminary image evaluation. *Journal of Medical Radiation Sciences* **2024**, *71*, 540 – 546. https://doi.org/10.1002/jmrs.809.
25. ROZWAG, C.; VALENTINI, F.; COTTEN, A.; DEMONDION, X.; PREUX, P.; JACQUES, T. Elbow trauma in children: development and evaluation of radiological artificial intelligence models. *Research in Diagnostic and Interventional Imaging* **2023**, *6*, 100029. https://doi.org/https://doi.org/10.1016/j.redii.2023.100029.
26. Huhtanen, J.; Nyman, M.; Doncenco, D.; et al. Deep learning accurately classifies elbow joint effusion in adult and pediatric radiographs. *Scientific Reports* **2022**, *12*. https://doi.org/10.1038/s41598-022-16154-x.

 doi:10.20944/preprints202511.0873.v2

27. Erzen, E.M.; BÜtÜn, E.; Al-Antari, M.A.; Saleh, R.A.A.; Addo, D. Artificial Intelligence Computer-Aided Diagnosis to automatically predict the Pediatric Wrist Trauma using Medical X-ray Images. In Proceedings of the 2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS), 2023, pp. 1–7. https://doi.org/10.1109/ISAS60782.2023.10391582.
28. Ferdi, A. Lightweight G-YOLOv11: Advancing Efficient Fracture Detection in Pediatric Wrist X-rays, 2024, [arXiv:eess.IV/2501.00647].
29. Altmann-Schneider, I.; Kellenberger, C.J.; Pistorius, S.M.; Saladin, C.; Schäfer, D.; Arslan, N.; Fischer, H.L.; Seiler, M. Artificial intelligence-based detection of paediatric appendicular skeletal fractures: performance and limitations for common fracture types and locations. *Pediatric Radiology* **2024**, *54*, 136–145.
30. Zech, J.R.; Ezuma, C.O.; Patel, S.; Edwards, C.R.; Posner, R.; Hannon, E.; Williams, F.; Lala, S.V.; Ahmad, Z.Y.; Moy, M.P.; et al. Artificial intelligence improves resident detection of pediatric and young adult upper extremity fractures. *Skeletal Radiology* **2024**, *53*, 2643–2651.
31. Khanapure, A.; Kashyap, H.; Bidargaddi, A.; Habib, S.; Anand, A.; M, M.S. Bone Fracture Detection with X-Ray images using MobileNet V3 Architecture. In Proceedings of the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), 2024, pp. 1–8. https://doi.org/10.1109/I2CT61223.2024.10544356.
32. Yadav, D.P.; Sharma, A.; Athithan, S.; Bhola, A.; Sharma, B.; Dhaou, I.B. Hybrid SFNet Model for Bone Fracture Detection and Classification Using ML/DL. *Sensors* **2022**, *22*. https://doi.org/10.3390/s22155823.
33. Varun, V.; Natarajan, S.K.; M, A.; P, N.; A, M.C.; Moorthi Hosahalli, N. Efficient CNN-Based Bone Fracture Detection in X-Ray Radiographs with MobileNetV2. In Proceedings of the 2024 2nd International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS), 2024, pp. 72–77. https://doi.org/10.1109/ICRAIS62903.2024.10811726.
34. Handoko, A.B.; Putra, V.C.; Setyawan, I.; Utomo, D.; Lee, J.; Timotius, I.K. Evaluation of YOLO-X and MobileNetV2 as Face Mask Detection Algorithms. In Proceedings of the 2022 IEEE Industrial Electronics and Applications Conference (IEACon), 2022, pp. 105–110. https://doi.org/10.1109/IEACon55029.2022.9951831.
35. Selcuk, B.; Serif, S.; Serif, T. A Comparative Study on Distal Radius Fracture Detection: YOLOv8 and YOLOv11 Versus Faster R-CNN. In Proceedings of the Mobile Web and Intelligent Information Systems; Younas, M.; Awan, I.; Martin, L.; Wu, H., Eds., Cham, 2026; pp. 229–242.
36. Chen, S.; Peng, Y.; Liu, Y.; Wang, P.; Liu, T. Enhancing YOLOv11 with Large Kernel Attention and Multi-Scale Fusion for Accurate Small and Multi-Lesion Bone Tumor Detection in Radiographs. *Diagnostics* **2025**, *15*, 1988.
37. Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J.W. A theory of learning from different domains. *Machine learning* **2010**, *79*, 151–175.
38. Liu, W.; Hu, J.; Lv, F.; Tang, Z. A new method for long-term temperature compensation of structural health monitoring by ultrasonic guided wave. *Measurement* **2025**, *252*, 117310. https://doi.org/https://doi.org/10.1016/j.measurement.2025.117310.